University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Journal of Actuarial Practice 1993-2006 | Finance Department

1998

Journal of Actuarial Practice, Volume 6, Nos. 1 and 2, 1998

Colin Ramsay , Editor
University of Nebraska - Lincoln, cramsay@unl.edu

Follow this and additional works at: http://digitalcommons.unl.edu/joap

Part of the Accounting Commons, Business Administration, Management, and Operations Commons, Corporate Finance Commons, Finance and Financial Management Commons, Insurance Commons, and the Management Sciences and Quantitative Methods Commons

Ramsay, Colin , Editor, "*Journal of Actuarial Practice,* Volume 6, Nos. 1 and 2, 1998" (1998). *Journal of Actuarial Practice 1993-2006*. 61.
http://digitalcommons.unl.edu/joap/61

This Article is brought to you for free and open access by the Finance Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Journal of Actuarial Practice 1993-2006 by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Journal of Actuarial Practice Volume 6, Nos. 1 and 2, 1998

ARTICLES

Journal of Actuarial Practice

EDITORIAL POLICY

The aim of this international journal is to publish articles pertaining to the "art" and/or "science" involved in contemporary actuarial practice.

The Journal welcomes articles providing new ideas, strategies, or techniques (or articles improving existing ones) that can be used by practicing actuaries. One of the goals of the Journal of Actuarial Practice is to improve communication between the practicing and academic actuarial communities. In addition, the Journal provides a forum for the presentation and discussion of ideas, issues (controversial or otherwise), and methods of interest to actuaries.

The Journal publishes articles in a wide variety of formats, including technical papers, commentaries/opinions, discussions, essays, book reviews, and letters. The technical papers published in the Journal are neither abstract nor esoteric; they are practical and readable. Topics suitable for this journal include the following:

AIDS	financial reporting	pricing issues
annuity products	group insurance	product development
asset-liability matching	health insurance	reinsurance
cash-flow testing	individual risk taking	reserving issues
casualty ratemaking	insurance regulations	risk-based capital
credibility theory	international issues	risk theory
credit insurance	investments	social insurance
disability insurance	liability insurance	solvency issues
expense analysis	loss reserves	taxation
experience studies	marketing	valuation issues
FASB issues	pensions	workers' compensation

REVIEW PROCESS

A paper submitted to the Journal first is screened by the editor for suitability. If it is deemed suitable, copies are sent to several independent referees. The name of the author(s) of the paper under consideration is usually anonymous to the referees, and the identities of referees are never revealed to the author(s).

The paper is reviewed for content, originality, and clarity of exposition. On the basis of the referee reports, the editor makes one of the following decisions: (1) accept subject to minor revisions, (2) accept subject to major revisions, or (3) reject.

The editor communicates the recommendation to the author(s) along with copies of the referees' reports. The entire process is expected to take three to four months.

See back cover for instructions to authors.

EDITOR

Colin Ramsay
University of Nebraska

ASSOCIATE EDITORS

Robert Brown
University of Waterloo

Cecil Bykerk
Mutual of Omaha

Ruy Cardoso
Ernst & Young LLP

Samuel Cox
Georgia State University

David Cummins
University of Pennsylvania

Robert Finger
Milliman & Robertson, Inc.

Charles Fuhrer
The Segal Company

Farrokh Guiahi
Hofstra University

Steven Haberman
City University

Merlin Jetton
Allstate Life

Eric Klieber
Buck Consultants

Edward Mailander
Ernst & Young LLP

Charles McClenahan
William M. Mercer, Inc.

Robert Myers
Temple University

Norman Nodulman
CNA Insurance

François Outreville
United Nations

Timothy Pfeifer
Milliman & Robertson, Inc.

Esther Portnoy
University of Illinois

Robert Reitano
John Hancock Mutual Life

Alice Rosenblatt
Coopers & Lybrand LLP

Arnold Shapiro
Penn State University

Elias Shiu
University of Iowa

Michael Sze
Hewitt Associates

Joseph Tan
National Actuarial Network

Ronnie Tan
AEGON

Richard Wendt
Tower Perrin

Colin Ramsay
Managing Editor

Margo Young
Technical Editor

© Copyright November 1998

Absalom Press, Inc.
P.O. Box 22098, Lincoln, NE 68542-2098, USA.

Journal of Actuarial Practice ISSN 1064-6647

Journal of Actuarial Practice Vol. 6, 1998

Principles and Application of Credibility Theory

Vincent Goulet*

Abstract†

We review the history of the practical development of credibility theory. Emphasis is placed on the two main approaches to credibility theory: limited fluctuation credibility and greatest accuracy credibility. We explain when each approach should and should not be used. The presentation of greatest accuracy credibility theory starts with a review of (exact) Bayesian credibility and then moves to the Bühlmann-Straub model. Estimators of the structure parameters are discussed. Examples are presented to illustrate the concepts. Finally, the hierarchical credibility and crossed classification credibility models are presented.

Key words and phrases: *experience rating, limited fluctuation, greatest accuracy, hierarchical, crossed classification, structure parameter, estimation.*

1 Introduction

1.1 Experience Rating

The first concern of an insurer when establishing a base premium is to ensure that the premium is sufficiently large to fulfill its obligations. Only then will the insurer seek to distribute premiums fairly among its

*Vincent Goulet, Ph.D., is an assistant professor of actuarial science at Concordia University, Canada. He received his Ph.D. in actuarial science at the University of Lausanne, Switzerland.

Dr. Goulet's address is: Concordia University, 7141 Sherbrooke West, Montreal PQ H4B 1R6, CANADA. Internet address: *vgoulet@alcor.concordia.ca*

†This paper is awarded the 1998 Actuarial "Art & Science" Education Contest prize. The research for this paper is supported by Quebec's FCAR fund. The author thanks Roger Goulet for his time and patience while correcting the grammar in the earlier drafts of this paper; Professor François Dufresne for his many fruitful comments; and the two anonymous referees and the editor for many suggestions that improved this paper.

insureds.[1] In lines of business where the number of policies is large enough to allow it, the development of a classification system is usually the first step to achieve a fair premium distribution. Experience rating systems in general and credibility theoretic methods in particular then constitute an efficient second step to determine a fair premium distribution.

As the name suggests, an experience rating system takes into account the past individual experience of an insured when establishing the insured's premium. As such, these systems have a somewhat limited scope in insurance because they require the accumulation of a significant volume of experience. Experience rating is especially suited to certain lines of insurance such as workers compensation and automobile insurance; it is not used, for example, in traditional individual life insurance (one only dies once) or homeowners insurance, where the claim frequency is low.

On a more formal basis, Bühlmann (1969) defines experience rating as follows:

Definition 1 (Experience Rating). *Experience rating aims at assigning to each individual risk its own* correct premium *(rate). The* correct premium *for any period depends exclusively on the (unknown) claims distribution of the individual risk for this same period.*

To illustrate and clarify the concept of experience rating, the following example (taken from Norberg (1979) with some modifications) is provided.

Example 1. Let us assume that a portfolio consists of ten insureds who are considered a priori to be equivalent on a risk level basis. Moreover, an insured can incur, at most, one claim per year, the severity of that claim being 1. The premium for this portfolio, called the *collective premium*, is estimated to be 0.20 and, accordingly, this is the premium every insured pays in the first year. After one year, the insurer observes the claim record shown in Figure 1 (where zeros corresponding to claim-free records are deleted to increase readability). The average claim amount is $1/10 = 0.10$. This is significantly below the assumed average of 0.20. Due to the limited experience in both the number of insureds and the number of years duration of the policy, however, the insurer is inclined to keep its premium unchanged.

After two years the average claim cost amounts to $4/20 = 0.20$; see Figure 2. Though the collective premium still seems adequate, one

[1] Throughout this paper the term *insured* is used in a broad sense. Depending on the line of business, an insured could be a person or a group of persons, a company, a reinsurance treaty, or any other adherent to an insurance contract.

Figure 1
Portfolio Experience After One Year

Year	Insureds									
	1	2	3	4	5	6	7	8	9	10
1									1	

notes that insured number 9 exhibits the worst record. Is this due only to bad luck? Unfortunately, due to the limited volume of experience, the insured cannot come to any conclusion on the general risk level of the portfolio or of any of the individual insureds.

Figure 2
Portfolio Experience After Two Years

Year	Insureds									
	1	2	3	4	5	6	7	8	9	10
1									1	
2	1	1							1	

Let us now jump eight years forward, at a time where the insurer is better able to infer results about the individual insureds' level of risk from the portfolio data. The data after ten years are depicted in Figure 3. One can see that the overall claim average, $\bar{X}$, is $23/100 = 0.23$. It is thus reasonable to think that the collective premium is adequate or even too low. The individual average for insured i, $\bar{X}_i$, on the other hand, shows great disparities among the insureds. In particular, the suspicion about insured 9 is confirmed: its 0.7 ratio suggests a risk worse than the collective one. Insureds 7, 8, and 10, however, incurred no claims. This ends the example.

If the collective premium in this example is *globally adequate*, it is in return clearly not *fair*. Some insureds deserve to pay a higher premium, while some should pay less. Though the portfolio was at first considered to be composed of more or less equivalent risks, experience has shown that the portfolio is, to some degree, heterogeneous. It is thus for equity concerns (and, perhaps, to gain a competitive edge) that insurers should, whenever possible, consider individual experience in ratemaking. In other words, the portfolio's heterogeneity forces the insurer to do experience rating.

Figure 3
Portfolio Experience After Ten Years

	Insureds									
Year	1	2	3	4	5	6	7	8	9	10
1									1	
2	1	1							1	
3	1		1						1	
4			1						1	
5									1	
6		1								
7	1	1		1	1					
8	1			1		1			1	
9	1				1					
10	1								1	
$\bar{X}_i$	0.6	0.3	0.2	0.2	0.2	0.1	0	0	0.7	0
$\bar{X}$	0.23									

1.2 An Overview of the Paper

There are many different experience rating systems, including bonus-malus systems, merit-demerit systems, participating policies, and commissions in reinsurance; see, for example, Bühlmann (1967, 1969). The most widely used methods, however, are based on credibility theory. Credibility theory uses two main approaches, each representing a different method of incorporating individual experience in the ratemaking process. The first and oldest approach is called *limited fluctuation* credibility (also referred to as *American credibility*). According to this approach, an insured's premium should be based solely on its own experience if the experience is significant and stable enough to be considered credible.

The second approach is called *greatest accuracy* credibility (also referred to as *European credibility*). It does not concentrate on the stability of the experience, but rather it focuses on the homogeneity of the experience within the portfolio. It would then be justifiable to give some weight to individual experience, provided it is significantly different from the portfolio's. The more heterogeneous the portfolio, the more important becomes individual experience and vice-versa.

This paper covers both the limited fluctuation and greatest accuracy approaches with the hope of clearing up the often blurred distinctions between them. Section 2 contains a brief discussion of the origins of limited fluctuation and greatest accuracy credibility theories. Section 3 describes the mathematical foundations of limited fluctuation credibility within the framework of the collective model of risk theory. The most important formulae are presented and illustrated in two examples. Some comments follow on the uses and misuses of the model in practice. The remainder of the paper is devoted to greatest accuracy credibility theory. Section 4 describes the mathematical foundations of greatest accuracy credibility theory within the framework of the collective model of risk theory. Section 5 presents exact Bayesian credibility theory, which is one approach used to determine the greatest accuracy credibility premium. The main results of exact Bayesian credibility are summarized in Tables 1 and 2. Section 6 is devoted to the well-known Bühlmann–Straub model. The credibility premium is presented and interpreted. Two useful generalizations of the Bühlmann–Straub model are introduced: the hierarchical credibility theory (Section 7) and crossed classification models (Section 8).

Finally, many of the theoretical results of credibility theory are described without any proofs or long mathematical developments. Emphasis is placed on the interpretation of results and discussion of practical issues. Advanced mathematical and technical expositions have been deliberately avoided; they can be found in many of the numerous suggested references listed at the end of this paper.

2 A Brief Historical Review

2.1 Limited Fluctuation Credibility Theory

The birth of credibility theory dates back to the beginning of the century with a paper by Mowbray (1914). In the workers compensation insurance field, Mowbray was interested in finding the minimal number of employees covered by a plan such that the premium of the employer could be considered fully dependable, that is, fully *credible*. Assuming that the probability of an accident, θ, is known, Mowbray wanted to calculate the minimum number of employees, n, so that the number of accidents would lie within $100k$ percent of the average (or mode) with probability p. If N denotes the total number of accidents of an employer, Mowbray's problem can be written as:

$$P[(1-k)\mathrm{E}[N] \leq N \leq (1+k)\mathrm{E}[N]] \geq p,$$

where $N \sim \text{Binomial}(n, \theta)$, i.e., N is binomial with mean $n\theta$ and variance $n\theta(1-\theta)$. Using the normal approximation for N eliminates the choice between the mean and the mode and yields:

$$n \geq \left(\frac{\zeta_{1-\varepsilon/2}}{k}\right)^2 \frac{(1-\theta)}{\theta} \tag{1}$$

where $\varepsilon = 1 - p$ and ζ_α is the αth percentile of a standard normal distribution.

Mowbray's solution needed only a distribution for N, the total number of claims, in order to determine a full credibility level. Unfortunately, however, his solution provided just that, a level above which an individual premium is granted *full* credibility and *zero* credibility below that level. Thus, an insured with total number of claims just below the full credibility level may pay a significantly different premium.[2]

The dichotomy between zero and full credibility paved the way for the development of *partial credibility*. The first formal theory was developed by Albert W. Whitney. In his 1918 paper, Whitney refers to "the necessity, from the standpoint of equity to the individual risk, of striking a balance between class-experience on the one hand and risk-experience on the other." The objective of credibility theory is the calculation of this balance.

Which principles should govern the calculation of this balance? According to Whitney (1918), the balance depends on four elements: the exposure, the hazard, the credibility of the manual rate (collective premium), and the degree of concentration within the class.[3] Moreover, Whitney (1918) writes:

> There would be no experience-rating problem if every risk within the class were typical of the class, for in that case the diversity in the experience would be purely adventitious.

Whitney's approach to the partial credibility problem is the first step toward greatest accuracy credibility, based on the homogeneity of the portfolio.

---

[2]In Mowbray's day some actuaries believed no data set was ever 100 percent reliable.

[3]The degree of concentration within the class is referred to as the *homogeneity* (similarity of individual experiences) of the entire portfolio.

Whitney's model for the homogeneity of the portfolio assumes that the individual averages are distributed according to a normal distribution. After some lengthy calculations, Whitney obtains the following expression for the individual's premium, P:

$$P = zX + (1 - z)m, \tag{2}$$

where X is the mean from the individual's experience and m the collective mean. Notice that X and m are combined to produce a weighted average with z and $1 - z$ as weights. An expression of the form of equation (2) is called a *credibility premium.* The quantity z is called the *credibility factor* and Whitney's expression is of the form

$$z = \frac{n}{n + K} \tag{3}$$

where K is a constant. Note that K is *not* an arbitrary constant, rather it is an explicit expression that depends on the various parameters of the model. For the sake of simplicity and to avoid large fluctuations between the individual and collective premiums, however, Whitney suggests that K be determined by the actuary's judgment rather than by its correct mathematical formula. Whitney's suggestion results in a stability-oriented form of credibility theory rather than a precision-based one. Thus, the birth of greatest accuracy credibility theory was delayed for almost half a century. Nevertheless, the determination of K by the actuary's judgment has since been widely and successfully used by American actuaries.

2.2 Greatest Accuracy Credibility Theory

One of the reasons why the greatest accuracy approach to credibility theory was slow to develop may already be found in a discussion of Whitney's paper. Fischer (1919) criticizes Whitney's use of the first version of Bayes' Rule where, a priori, all possible events are equally likely to occur. This rule was called the "principle of insufficient reason" by its proponents while its detractors called it the "assumption of the equal distribution of ignorance." In addition, until the mid 1950s, there was a general negative attitude in the American statistical community toward what is known today as neo-Bayesian statistics.

Greatest accuracy credibility theory originated from two seminal papers by Bailey (1945, 1950). In his 1945 paper, Bailey obtains a credibility formula that seems to anticipate the nonparametric universe to be explored two decades later by Bühlmann. Unfortunately, the paper suffered due to a somewhat awkward notation that made it difficult to read. The 1950 paper, on the other hand, was better understood and is considered as the pioneering paper in greatest accuracy credibility.

Bailey (1950) writes:

> At present, practically all methods of statistical estimation appearing in textbooks on statistical methods or taught in American universities are based on an equivalent to the assumption that any and all collateral information or a priori knowledge is worthless. ... Philosophers have recently discussed the credibilities to be given to various elements of knowledge (Russell 1948), thus undermining the accepted philosophy of the statisticians. However, it appears to be only in the actuarial field that there has been an organized revolt against discarding all prior knowledge when an estimate is to be made using newly acquired data.

Here Bailey is advocating the Bayesian philosophy with the proviso that Laplace's generalization of Bayes' rule be used instead of the original Bayes rule. With this generalization, the Bayes' rule is applicable even if possible events have varying probabilities of occurring.

Bailey then shows that the Bayesian estimator obtained by minimizing the mean square error is a linear function of the observations, corresponding exactly with the credibility premium for the combinations of conjugate prior distributions such as binomial/beta, Poisson/gamma, and normal/normal. He is among the first to discover this linearity of the Bayesian estimator.[4] His credibility factor is still of the form $z = n/(n + K)$, where K depends on the parameters of the model. Unlike Whitney, however, Bailey does not propose to evaluate K using the actuary's judgment, but rather sticks to its algebraic expression.

Meanwhile, a new branch of statistics called *empirical Bayesian* statistics was being developed by Robbins (1955, 1964). It would be of importance in the future development of credibility theory because it filled the gap between theory and practice. One of the main problems with the Bayesian approach is the need to know the prior distribution, a condition seldom met in practice. Robbins' empirical Bayes approach is to assume that, although unknown, the prior distribution does exist and can be estimated from repeated similar experiences. Robbins (1964) writes:

---

[4]Norberg (1979) states that Keffer (1929) obtained a similar result in the Poisson/gamma case and that there would exist some earlier references.

> The empirical Bayes approach to statistical decision problems is applicable when the same decision problem presents itself repeatedly and independently with a fixed unknown a priori distribution of the parameter.

As Bühlmann points out later, this applies perfectly to the experience rating problem.

Given the fact that actuaries wish to have linear credibility premiums (as linear premiums are easy to calculate and easy to explain), Bühlmann suggested at the 1965 ASTIN Colloquium in Lucerne, Switzerland, that the Bayesian estimator be forced to be a linear combination of the observations. In a nonparametric setting, Bühlmann (1967, 1969) derives a linear expression featuring a credibility factor of the well-known form $z = n/(n + K)$, with a simple and general expression for the constant K.

The 1970s heralded the rapid development of credibility theory. Bühlmann and Straub (1970) generalize Bühlmann's classical model by assigning weights to the observations and by introducing estimators for the structure parameters.[5] This was followed by two important generalizations of the Bühlmann and Bühlmann-Straub models: the hierarchical model due to Jewell (1975) and the linear regression model due to Hachemeister (1975). The following year, De Vylder (1976b) presented a semilinear and an optimal semilinear credibility model together with the first formulation of the credibility problem in terms of Hilbert spaces (De Vylder 1976a). Three years later, Norberg (1979) published an extensive paper in which he reviewed most of what was known in credibility theory. This paper still remains a key reference in credibility theory.

While in the 1970s the bulk of the credibility research was focused on model generalizations, during the 1980s research was focused primarily on the estimation of structure parameters. The important papers include De Vylder (1978, 1981, 1984), Norberg (1980), Gisler (1980) and Dubey and Gisler (1981). From the mid-1980s to the early 1990s, research in credibility theory slowed until a revival of interest stimulated by optimal parameter estimation (De Vylder and Goovaerts 1991, 1992*a*, 1992*b*) and robust parameter estimation (Künsch 1992, Gisler and Reinhard 1993).

A recent innovation in credibility theory is the variance components model introduced by Dannenburg (1995) to describe his crossed classification credibility model. This is briefly studied in Section 8.

---

[5]These improvements led to a wider use of greatest accuracy credibility in practice, although mostly in Europe.

3 Limited Fluctuation Credibility

Limited fluctuation credibility originated in the early 1900s with Mowbray's paper "How Extensive A Payroll Exposure Is Necessary To Give A Dependable Pure Premium?" As the title states, Mowbray was interested in finding the level of payroll in workers compensation insurance for which the pure premium of a given insured can be considered fully credible.

An individual insured's experience is considered to be *fully credible* if it fluctuates moderately from one period to another. That is, the credibility criterion is stability of experience, which usually increases with the volume of the insured's experience. This volume can be expressed as premium volume, number of claims, number of employees, square foot of factory surface, etc.

3.1 The General Model

With the emergence of risk theoretic methods, Mowbray's original problem can be formulated in a slightly more general way as follows. Let us define the random variables

N_t = The number of claims the insured generated during the tth time period (months, quarters, years, etc.), for $t = 1, 2, \ldots$;

X_{tj} = Size of the jth claim in the tth year, for $j = 1, 2, \ldots, N_t$;

S_t = The size of the aggregate claims in the tth period of time.

Then,

$$S_t = X_{t1} + X_{t2} + \cdots + X_{tN_t} \tag{4}$$

where X_{tj}s are assumed to be independent, identically distributed (i.i.d.) random variables that are also mutually independent of the N_ts. This is the classical collective model of risk theory. Most of the situations usually encountered in limited fluctuation credibility can be described by an application of this model. It is also well-known (see, for example, Gerber (1979)) that

$$\mathrm{E}[S_t] = \mathrm{E}[N_t]\,\mathrm{E}\left[X_{tj}\right] \tag{5}$$

and

$$\mathrm{Var}[S_t] \quad = \quad \mathrm{E}[N_t]\,\mathrm{Var}\left[X_{tj}\right] + \mathrm{Var}[N_t]\,\mathrm{E}\left[X_{tj}\right]^2. \tag{6}$$

Let $\bar{S}_T = (S_1 + S_2 + \cdots + S_T)/T$ denote the insured's observed average (empirical mean) claim amount at the end of T periods, $T = 1, 2, \ldots$. The fundamental problem of limited fluctuation credibility is the determination of the parameters of the distribution of $\bar{S}_T$ such that it stays within $100k$ percent of its expected value with probability p, i.e.,

$$P\left[(1-k)\mathrm{E}\left[\bar{S}_T\right] \leq \bar{S}_T \leq (1+k)\mathrm{E}\left[\bar{S}_T\right]\right] \geq p, \tag{7}$$

holds for given p and k. In a typical limited fluctuation credibility situation, the parameter k is small (e.g., 5 to 10 percent), while parameter p is large (often above 90 percent).

When an insured meets the requirements of equation (7), the insured is said to deserve a full credibility of order (k, p), i.e., the insured is charged a pure premium based solely on the insured's own claims experience. If full credibility occurs after T^* periods the credibility premium would be $\bar{S}_{T^*}$.

Equation (7) thus requires the distribution of $\bar{S}_T$ to be relatively concentrated around its mean. As $\bar{S}_T$ is a sum of i.i.d. random variables, the distribution of $\bar{S}_T$ has to be approximated. Assuming the second moment of $\bar{S}_T$ is finite, one can use the version of the central limit theorem applicable to random sums (Feller 1966, p. 258) to approximate the distribution. Thus,

$$\frac{(\bar{S}_T - \mathrm{E}\left[\bar{S}_T\right])}{\sqrt{\mathrm{Var}\left[\bar{S}_T\right]}} \sim N(0,1),$$

i.e., a standard normal distribution. Equation (7) may then be rewritten

$$\Pr\left[\left|\frac{(\bar{S}_T - \mathrm{E}\left[\bar{S}_T\right])}{\sqrt{\mathrm{Var}\left[\bar{S}_T\right]}}\right| \leq \frac{k\mathrm{E}\left[\bar{S}_T\right]}{\sqrt{\mathrm{Var}\left[\bar{S}_T\right]}}\right] \approx 2\,\Phi\left(\frac{k\mathrm{E}\left[\bar{S}_T\right]}{\sqrt{\mathrm{Var}\left[\bar{S}_T\right]}}\right) - 1 \geq p, \tag{8}$$

hence

$$\left(\mathrm{E}\left[\bar{S}_T\right]\right)^2 \geq \left(\frac{\zeta_{1-\varepsilon/2}}{k}\right)^2 \frac{\mathrm{Var}[S_t]}{T}, \tag{9}$$

where $\varepsilon = 1 - p$ and ζ_α is the αth percentile of a standard normal distribution.

At this point, the essence of the theory of limited fluctuation credibility (i.e., equation (7)) has been covered. What follows are examples of the calculations needed to satisfy equation (7). These calculations are more relevant to general risk theory, however, than to credibility theory.

Example 2. Recall the assumptions of Example 1 above. In that example, an insured can incur at most one claim per year, the severity of that claim being 1. Thus the distribution of S_t is Bernoulli with parameter θ, i.e., $\Pr[S_t = 1] = \theta$ and $\Pr[S_t = 0] = 1 - \theta$. Thus $\mathrm{E}\left[\bar{S}_T\right] = \theta$ and $\mathrm{Var}\left[\bar{S}_T\right] = \theta(1-\theta)/T$. From equation (7), the full credibility level of order (k, p) is given by:

$$T \geq \left(\frac{\zeta_{1-\varepsilon/2}}{k}\right)^2 \frac{1-\theta}{\theta}.$$

If we further assume that $\theta = 0.20$, $k = 0.05$ and $p = 0.90$ then the full credibility level of order $(0.05, 0.90)$ occurs after $T = 4323$ years of experience!

Example 3. Suppose the insured can incur at most n claims per year, the severity of each claim being 1. The claims are assumed to occur independently with probability θ per occurrence. Thus the distribution of S_t is binomial(n, θ). The full credibility level of order (k, p) is given by:

$$nT \geq \left(\frac{z_{1-\varepsilon/2}}{k}\right)^2 \frac{1-\theta}{\theta}.$$

As expected, there is an inverse relationship between n and T. Thus if, for example, the expected annual aggregate claims $n\theta$ is small, then we need more years for a credible claims history to develop, i.e., larger T.

Example 4. The most widely used distribution for S_t is the one where N_t has a Poisson distribution with parameter λ giving S_t a compound Poisson distribution. From equations (5) and (6), $\mathrm{E}[S_t] = \lambda \mathrm{E}\left[X_j\right]$ and $\mathrm{Var}[S_t] = \lambda \mathrm{E}\left[X_j^2\right]$. The full credibility level of order (k, p) is thus given by:

$$T\lambda \geq \left(\frac{z_{1-\varepsilon/2}}{k}\right)^2 \left[1 + \frac{\mathrm{Var}\left[X_j\right]}{\mathrm{E}\left[X_j\right]^2}\right].$$

Again, there is an inverse relationship this time between λ and T. Thus if, for example, the expected annual number claims λ is small, then we need more years for a credible claims history to develop, i.e., larger T. Note that the choice $k = 5$ percent, $p = 0.90$, and $\Pr[X_j = 1] = 1$ leads to the famous λ value of 1,082.

One may also like to refer to Longley-Cook (1962) for some more examples involving limited fluctuation credibility.

3.2 Using Other Approximation Methods

In general, the distribution of S_t is not symmetrical, even if that of X_j is. A normal approximation is nevertheless used to calculate the full credibility levels because, as seen in equation (9), it easily leads to simple formulae.

One might wonder if using more refined approximations taking the skewness of S_t into account would lead to better or more accurate full credibility levels. Normal power and Esscher approximations are two examples that account for the skewness of S. Goulet (1997) shows, however, that the effect of using these approximation methods is negligible in almost any case. Thus, more sophisticated approximation methods are not worth the added complexity and calculation time when compared to the normal approximation.

3.3 Partial Credibilities

As mentioned in Section 2, the first partial credibility formula is due to Whitney (1918), who was motivated by his desire to obtain a premium that struck a balance between the individual premium of a single insured and the manual or collective premium of the entire portfolio to which the insured belongs.

Since 1918, many partial credibility formulae have been proposed. Among the three most widely used are:

$$\begin{aligned} z_1 &= \min\left\{\sqrt{\frac{n}{n_0}}, 1\right\}, \\ z_2 &= \min\left\{\left(\frac{n}{n_0}\right)^{2/3}, 1\right\}, \end{aligned}$$

and

$$z_3 = \frac{n}{n+K},$$

where n_0 is the full credibility level and K a constant determined by the actuary's judgment. One consideration in the choice of K is the desire to limit size of the changes in the premium from one year to the next. The third partial credibility formula, z_3, is the one proposed by Whitney. In addition z_3 is the only one in which the (partial) credibility level never reaches unity.

3.4 Uses of Limited Fluctuation Credibility

From a theoretical perspective, the range of applications of limited fluctuation credibility is fairly limited, though many of these are ignored in practice. The key point to remember when using limited fluctuation credibility is that it relies solely on a stability criterion, which, generally, is the size of the insureds or the number of periods (years quarters, etc.) of claims experience. As such, limited fluctuation credibility should be used only when stability of the experience is of foremost importance. One good example is the determination of an admissibility threshold in a retrospective insurance system, where the insured's premium is readjusted at the end of the year after the total claim amount is known.

The case for partial credibility is even more delicate. Since its inception, partial credibility has been successfully used by American actuaries to restrict premium variation from one time period to another. One can argue that partial credibility takes into account the heterogeneity of the insurer's block of insureds by charging different premiums to different groups of insureds. This differentiation among the insureds, however, is only based on their size or the extent of their claims history; this is not necessarily fair.

One must bear in mind that the goal of partial limited fluctuation credibility is not to calculate the most precise premium for an insured. The goal is to incorporate into the premium as much individual experience as possible while still keeping the premium sufficiently stable. It is important to understand this distinction. When credibility is used to find the most precise estimate of an insured's pure risk premium, one must turn to greatest accuracy credibility methods.

The remainder of this paper is devoted to the various forms of greatest accuracy credibility theory.

4 Greatest Accuracy Credibility: An Overview

Greatest accuracy credibility is a more modern, versatile, and complex field of credibility theory. It is not a single theory; rather it is an approach to the credibility problem. The approach is to find the best premium to charge an insured, where best is in the sense that the premium estimator is the closest estimator to the true premium. The traditional starting point in the study of greatest accuracy credibility theory is Bayesian credibility theory, where the fundamental concepts can be illuminated in a parametric setting.[6]

One important point to keep in mind when moving from limited fluctuation to greatest accuracy credibility is that a high credibility factor (i.e., z close to one) is no longer a goal in itself. Indeed, the credibility factor will henceforth mostly reflect the degree of heterogeneity of the portfolio, rather than the degree of stability of an individual risk's experience. For a homogeneous portfolio, greatest accuracy credibility states there is no need to charge a different premium to the insureds. The credibility factor will accordingly be low, i.e., close to zero. Conversely, the more heterogeneous the portfolio, the greater the consideration of the individual experience; hence the higher the credibility factor.

To illustrate this, imagine a portfolio consisting of five very large insureds, each having identical means. Given the importance of their size, each group of insureds would all be granted full credibility under the limited fluctuation approach. As their means are all equal, however, they form a perfectly homogeneous portfolio. Accordingly, their credibility level will be zero under the greatest accuracy approach. Of course, the end result is the same because the collective mean is equal to the individual means, but this shows how different can be the interpretation of the credibility factor in greatest accuracy credibility.

4.1 The Mathematical Model

Consider an insurance portfolio consisting of I insureds. The ideal situation for ratemaking occurs if this portfolio is relatively homogeneous, i.e., the insureds have similar risk characteristics. The group of characteristics of insured i that reflects the insured's risk level is donated by the risk parameter Θ_i for insured $i = 1, \dots, I$. This risk parameter incorporates every characteristic of the insured that is not otherwise accounted for in the initial risk classification process.

The parameter Θ_i is unknown and is assumed to be constant throughout the life of the insurance contract. Because of the assumption of a homogeneous portfolio, we must further assume that each insured's Θ_i

[6]Norberg (1979) and Goovaerts and Hoogstad (1987) are also good references for those who would like to delve deeper into the subject.

is viewed as being drawn at random from the same cumulative distribution, $U(\theta)$. Following Bühlmann (1969), $U(\theta)$ is called the *structure function.* This is essentially an empirical Bayes approach where the structure function exists but is unknown and has to be estimated from the portfolio data.

In a purely Bayesian setting, $U(\theta)$ represents the insurer's prior belief about the insured's risk level. After collection of the insured's data at the end of the period, the insurer's initial judgment is revised and the structure function modified accordingly. This interpretation is particularly suited to the case where there is a single insured or when the insurer has little information and must make an educated guess at the initial pure premium—for example, when the insurer is entering a new line of business where no data are available.

Throughout the rest of this section, we consider the purely Bayesian setting with only one insured (so the subscript i will be dropped). The claim amounts X_t $(t = 1, 2, \ldots,)$ are independent and identically distributed, but only given θ, the risk parameter of the insured. Unconditionally, the X_ts are not necessarily independent. The conditional distribution of $X|\Theta = \theta$ is denoted by $F(x|\theta)$. The unconditional (or marginal) distribution function of X_t is given by

$$F(x) = \int_\theta F(x|\theta)\, dU(\theta). \tag{10}$$

The determination of a claim amount can thus be seen as a two-stage process: first obtain a risk level for the insured from the distribution $U(\cdot)$ and then a claim amount from the conditional distribution $F(\cdot|\theta)$. This two-stage model is also called an *urn of urn* model.

The two-stage process gives rise to the so-called apparent contagion phenomenon studied by Feller (1943). To illustrate this phenomenon, consider an insured chosen randomly from a homogeneous portfolio. Nothing is known about the insured except that the portfolio mean claim amount is, say, \$100. The insured's claim record observed during five years is as follows: 65, 72, 88, 69, and 76. These claim amounts are smaller than the portfolio average and thus seem to be positively correlated. If, on the other hand, the insured was known to have a mean claim of, say, \$75, then the observed claim amounts would simply appear as random and uncorrelated variations around this mean! The apparent dependency of the (unconditional) claim amounts X_t is only a consequence of the urn of urn sampling method. Successive claim amounts are, in reality, independent.

4.2 The Definitions of the Various Premiums

An underlying tenet of credibility theory is that the premium sought (estimated) is the pure or net premium, without any provision for random fluctuations, profits, or expenses. Thus two insureds with different variances but the same mean are charged the same pure premium. We distinguish here between four types of (pure) premiums: the risk premium, the collective premium, the Bayesian premium, and the credibility premium.

Definition 2 (Risk Premium). *The risk premium, $\mu(\theta)$, is the correct premium to charge an insured if the insured's risk level, θ, is known. The risk premium is thus the expected value of the insured's aggregate claim amount in one period, given his or her risk level.*

The risk premium is given by

$$\mu(\theta) = \mathrm{E}[X|\Theta = \theta] = \int_0^\infty x f(x|\theta)\,dx. \tag{11}$$

Because the risk parameter Θ is unobservable in practice, the risk premium can never be exactly known and hence must be estimated from data. At the other extreme is the collective premium.

Definition 3 (Collective Premium). *The collective premium, m, is the pure premium charged when nothing is known about the insured's risk level (during the first year, for example). The collective premium is in essence the average value of all possible risk premiums.*

Mathematically, the collective premium is given by

$$m = \mathrm{E}[X] = \mathrm{E}[\mathrm{E}[X|\Theta]] = \mathrm{E}[\mu(\Theta)]. \tag{12}$$

The fundamental difference between limited fluctuation and greatest accuracy credibility is the type of estimator for the risk premium. In limited fluctuation credibility, the observed claim average $\bar{X}$ is chosen if the experience is sufficiently stable and fully credible; otherwise the collective mean m is charged. In greatest accuracy credibility, on the other hand, the objective is to find an estimator as close as possible to the true value of $\mu(\theta)$ given the available data. There is no unique way to measure closeness. In Bayesian credibility, for example, the closeness measure is the mean square error between the estimator and the risk premium.

Definition 4 (Bayesian Premium). *Suppose the data for T consecutive periods are $X_1, \ldots, X_T$, then the Bayesian premium $\mathcal{B}(X_1, \ldots, X_T)$ is given by*

$$\mathcal{B}(X_1, \ldots, X_T) = \min_{g(\cdot)} \mathrm{E}\Big[(\mu(\Theta) - g(X_1, \ldots, X_T)^2)\Big], \tag{13}$$

where $g(\cdot)$ is some function of the data.

It is not difficult to prove (see, for example, Hogg and Craig (1978), Goovaerts and Hoogstad (1987)) that the solution to this minimization problem is

$$\mathcal{B}(X_1, \ldots, X_T) = \mathrm{E}[\mu(\Theta) | X_1, \ldots, X_T]. \tag{14}$$

Moreover, because the distribution of X_{T+1} (the next period's claim record) is identical to that of X_t, for $t = 1, 2, \ldots, T$, we also have

$$\mu(\Theta) = \mathrm{E}[X_{T+1} | \Theta] = \mathrm{E}[X_{T+1} | \Theta, X_1, \ldots, X_T]. \tag{15}$$

Therefore, the Bayesian premium can also be written as

$$\begin{aligned} \mathcal{B}(X_1, \ldots, X_T) &= \mathrm{E}[\mathrm{E}[X_{T+1} | \Theta, X_1, \ldots, X_T] \, | X_1, \ldots, X_T] \\ &= \mathrm{E}[X_{T+1} | X_1, \ldots, X_T]. \end{aligned} \tag{16}$$

This last expected value minimizes $\mathrm{E}[(X_{T+1} - g(X_1, \ldots, X_T))^2]$.

The Bayesian premium can thus be calculated in two different ways:

1. Directly from the *posterior* distribution of X_{T+1} given the data $X_1, \ldots, X_T$:

$$\mathrm{E}[X_{T+1} | X_1, \ldots, X_T] = \int_0^\infty x dF(x | x_1, \ldots, x_T); \tag{17}$$

2. Or in two steps by calculating first the posterior distribution of Θ given the data, $U(\theta | x_1, \ldots, x_T)$, and then by calculating the expected value of $\mu(\theta)$ with respect to this distribution:

$$\mathrm{E}[\mu(\Theta) | X_1, \ldots, X_T] = \int_\theta \mu(\theta) dU(\theta | x_1, \ldots, x_T). \tag{18}$$

Recall that the (conditional) distribution of $X_t|\Theta$ and the distribution of Θ are assumed to be known in the present model. Calculating the Bayesian premium by equation (17) first requires determination of the posterior distribution of X_{T+1}. By general properties of conditional and multivariate distributions (see, for example, Hogg and Craig (1978)) and by the conditional independence of claim amounts, we have

$$\begin{aligned} dF(x_{T+1}|x_1,\ldots,x_T) &= \frac{\int_\theta dF(x_{T+1},x_1,\ldots,x_T,\theta)\,d\theta}{\int_\theta dF(x_1,\ldots,x_T,\theta)\,d\theta} \\ &= \frac{\int_\theta dF(x_{T+1}|\theta)dF(x_1,\ldots,x_T|\theta)dU(\theta)}{\int_\theta dF(x_1,\ldots,x_T,\theta)\,d\theta} \\ &= \int_\theta dF(x_{T+1}|\theta)dU(\theta|x_1,\ldots,x_T). \end{aligned} \tag{19}$$

Using equation (18) requires the posterior distribution of Θ, but this calculation is usually easier than the preceding one. From Bayes theorem and the conditional independence of claim amounts,

$$\begin{aligned} dU(\theta|x_1,\ldots,x_T) &= \frac{dF(x_1,\ldots,x_n|\theta)dU(\theta)}{\int_\theta dF(x_1,\ldots,x_n|\theta)\,d\theta} \\ &= \frac{\prod_{j=1}^T dF(x_j|\theta)dU(\theta)}{\int_\theta \prod_{j=1}^T dF(x_j|\theta)\,d\theta} \\ &\propto dU(\theta)\prod_{j=1}^T dF(x_j|\theta). \end{aligned} \tag{20}$$

Calculation of the expected value is then immediate. Examples of such calculations are given in Section 5.

The last premium to define before we turn to exact Bayesian credibility is the credibility premium.

Definition 5 (Credibility Premium). *A credibility premium, $\mathcal{P}$ is a* linear *function of a special type of observations $X_1,\ldots,X_T$ of an insured: it is a convex combination of the individual experience weighted average ($\bar{X}$) and the collective premium (m), i.e.,*

$$\mathcal{P}(X_1,\ldots,X_T) = z\bar{X} + (1-z)m, \tag{21}$$

where $0 \le z \le 1$ is the credibility factor *and $(1-z)$ is the complement of credibility.*

It should be noted that the complement of credibility is given to the collective premium, m, and nothing else.

5 Exact Bayesian Credibility

To an actuary who considers himself or herself to be a Bayesian, the Bayesian premium equation (14) is the best premium (in the least square sense) to charge an insured considering the experience at hand. The Bayesian premium, however, has some drawbacks when it comes to being used in practice: the actual distributions of $X_t|\Theta$ and Θ must be known.

Moreover, unlike a credibility premium, there is no guarantee that a Bayesian premium will lie between the individual experience average $\bar{X}$ and the collective premium m. This fact can be difficult to explain to a layperson.

In some cases, the Bayesian premium can be extremely complicated. Fortunately, there are some combinations of distributions where the Bayesian premium has a nice form. Actually, in these cases, Bayesian premiums are exact credibility premiums.

Without loss of generality, the time periods in these examples are measured in years.

Example 5. Assume the probability of a claim of amount 1 occurring in year t is θ. Then the distribution of $X_t|\Theta = \theta$ is Bernouilli with parameter θ. The risk premium is $\mu(\Theta) = \Theta$ and, consequently, the collective premium is $m = \mathrm{E}[\Theta]$. If the distribution of Θ is uniform on (a, b), then Norberg (1979) shows that the Bayesian premium is

$$\mathcal{B}(X_1, \ldots, X_T) = \frac{\sum_{j=1}^{T-T\hat{\theta}} (-1)^j \frac{b^{T\hat{\theta}+j+2} - a^{T\hat{\theta}+j+2}}{(T-T\hat{\theta}-j)!\, j!\, (T\hat{\theta}+j+2)}}{\sum_{j=1}^{T-T\hat{\theta}} (-1)^j \frac{b^{T\hat{\theta}+j+1} - a^{T\hat{\theta}+j+1}}{(T-T\hat{\theta}-j)!\, j!\, (T\hat{\theta}+j+1)}},$$

where T is the number of years of experience and $\hat{\theta}$ the proportion of years where a claim has occurred.

Example 6. Here the distribution of Θ in the previous example is changed from a uniform distribution to a beta distribution with parameters α and β (see, for example, Hogg and Craig (1978)), i.e.,

$$dU(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 < \theta < 1,\ \alpha > 0,\ \beta > 0.$$

The expressions for the various quantities are easier to derive. The collective premium is

$$m = \mathrm{E}[\Theta] = \frac{\alpha}{\alpha + \beta}.$$

By equation (20) the posterior distribution of Θ with T years of experience is

$$\begin{aligned} dU(\theta|x_1,\ldots,x_T) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\prod_{j=1}^{T}\theta^{x_j}(1-\theta)^{1-x_j} \\ &\propto \theta^{\alpha+x-1}(1-\theta)^{\beta+T-x-1}, \end{aligned}$$

where $x = x_1 + \cdots + x_T$. By inspection, the posterior distribution of Θ is still beta, but with updated parameters $\tilde{\alpha} = \alpha + x$ and $\tilde{\beta} = \beta + T - x$. The Bayesian premium is therefore easily calculated as

$$\begin{aligned} \mathcal{B}(X_1,\ldots,X_T) &= \mathrm{E}[\Theta|X_1,\ldots,X_T] \\ &= \frac{\tilde{\alpha}}{\tilde{\alpha}+\tilde{\beta}} \\ &= \frac{\alpha + X_1 + \cdots + X_T}{\alpha+\beta+T} \\ &= z\bar{X} + (1-z)m, \end{aligned}$$

with $z = T/(T + \alpha + \beta)$. The Bayesian premium for the binomial/beta combination of distributions is thus a credibility premium with a credibility factor of the form $T/(T + K)$. Considering that T is the number of years, this is a credibility factor of the form $n/(n + K)$ that was discussed in Section 3.

Example 7. Suppose $X|\Theta$ has a Poisson distribution with parameter Θ, and Θ has a gamma distribution of parameters α and λ, i.e.,

$$dF(x|\theta) = \frac{\theta^x e^{-x}}{x!}, \quad x = 0, 1, \ldots$$

and

$$dU(\theta) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\lambda\theta}, \quad \theta > 0,\ \alpha > 0,\ \lambda > 0.$$

The risk premium is $\mu(\Theta) = \Theta$ and, consequently, the collective premium is $m = \mathrm{E}[\Theta] = \alpha/\lambda$. As in the preceding example, one finds that the posterior distribution of Θ is:

$$
\begin{aligned}
u(\theta|x_1,\ldots,x_T) &\propto \theta^{\alpha-1}e^{-\lambda\theta}\prod_{j=1}^{T}\theta^{x_j}e^{-\theta} \\
&\propto \theta^{\alpha+x-1}e^{(\lambda+T)\theta},
\end{aligned}
$$

which is gamma with updated parameters $\tilde{\alpha} = \alpha + x$ and $\tilde{\lambda} = \lambda + T$, where $x = x_1 + \cdots + x_T$. The Bayesian premium is thus

$$
\begin{aligned}
\mathcal{B}(X_1,\ldots,X_T) &= \frac{\tilde{\alpha}}{\tilde{\lambda}} \\
&= \frac{\alpha + X_1 + \cdots + X_T}{\lambda + T} \\
&= z\bar{X} + (1-z)m,
\end{aligned}
\tag{22}
$$

with $z = T/(T+\lambda)$. Once again, the Bayesian premium is a convex combination between the individual experience average and the collective premium, i.e., the credibility premium with credibility factor z.

As mentioned in Section 2, Bailey (1950) was one of the first to show that for some combinations of distributions the Bayesian estimator is exactly a (linear) credibility premium. In doing so, Bailey also provided the exact value of the constant K in the credibility factor that Whitney (1918) chose to determine by judgment. A few years later Mayerson (1964) extended Bailey's results.

All the combinations of distributions known to yield exact credibility premiums are presented in Tables 1 and 2.

Table 1
Bayesian Credibility Results for Certain Conjugate Distribution Pairs (Part 1)

	Conjugate Distribution Pairs		
	Bernouilli and Beta	Geometric and Beta	Normal and Normal
$dF(x\|\theta) =$	Bernouilli with $0 \le \theta \le 1$ $\theta^x(1-\theta)^{1-x}$ for $x = 0, 1$	Geometric with $0 \le \theta \le 1$ $\theta(1-\theta)^x$ for $x = 0, 1, \ldots$	$N(\theta, \sigma_2^2)$, $\sigma_2 > 0$ $\phi\left(\frac{x-\theta}{\sigma_2}\right)$ for $-\infty < x, \theta < \infty$
$dU(\theta) =$	Beta with α and β, $\alpha, \beta > 0$ $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$	Beta with α and β, $\alpha, \beta > 0$ $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$	$N(\theta, \sigma_1^2)$, $-\infty < \mu < \infty$ and $\sigma_1 > 0$ $\phi\left(\frac{\theta-\mu}{\sigma_1}\right)$
$dF(x) =$	$\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+x)\Gamma(\beta+1-x)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+1)}$	$\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+1)\Gamma(\beta+x)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+x+1)}$	$\phi\left(\frac{x-\mu}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)$
$dU(\theta\|x_1, \ldots, x_T) =$	Beta with $\tilde{\alpha}$ and $\tilde{\beta}$ $\frac{\Gamma(\tilde{\alpha}+\tilde{\beta})}{\Gamma(\tilde{\alpha})\Gamma(\tilde{\beta})}\theta^{\tilde{\alpha}-1}(1-\theta)^{\tilde{\beta}-1}$ where $\tilde{\alpha} = \alpha + \sum_j x_j$ and $\tilde{\beta} = \beta + T - \sum_j x_j$	Beta with $\tilde{\alpha}$ and $\tilde{\beta}$ $\frac{\Gamma(\tilde{\alpha}+\tilde{\beta})}{\Gamma(\tilde{\alpha})\Gamma(\tilde{\beta})}\theta^{\tilde{\alpha}-1}(1-\theta)^{\tilde{\beta}-1}$ where $\tilde{\alpha} = \alpha + T$ and $\tilde{\beta} = \beta + \sum_j x_j$	$N(\tilde{\mu}, \tilde{\sigma}_1^2)$ $\phi\left(\frac{x-\tilde{\mu}}{\tilde{\sigma}_1}\right)$ where $\tilde{\mu} = \frac{\sigma_1^2\sum_j x_j + \sigma_2^2\mu}{T\sigma_1^2+\sigma_2^2}$ and $\tilde{\sigma}_1^2 = \frac{\sigma_1^2\sigma_2^2}{T\sigma_1^2+\sigma_2^2}$
$\mu(\theta) =$	θ	$(1-\theta)/\theta$	θ
$m =$	$\alpha/(\alpha+\beta)$	$\beta/(\alpha-1)$	μ
$\mathcal{B}(X_1, \ldots, X_T) =$	$(\alpha + \sum_j X_j)/(\alpha+\beta+T)$	$(\beta + \sum_j X_j)/(\alpha+T-1)$	$(\sigma_1^2\sum_j X_j + \sigma_2^2\mu)/(T\sigma_1^2+\sigma_2^2)$
$z =$	$T/(T+\alpha+\beta)$	$T/(T+\alpha-1)$	$T/(T+\sigma_2^2/\sigma_1^2)$

Table 2
Bayesian Credibility Results for Certain Conjugate Distribution Pairs (Part 2)

	Conjugate Distribution Pairs	
	Poisson and Gamma	Exponential and Gamma
$dF(x\|\theta) =$	Poisson with $\theta > 0$ $\dfrac{\theta^x e^{-\theta}}{x!}$ for $x = 0, 1, \ldots$	Exponential with $\theta > 0$ $\theta e^{-\theta x}$ for $x > 0$
$dU(\theta) =$	Gamma with $\alpha, \lambda > 0$ $\dfrac{\lambda^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\lambda x}$	Gamma with $\alpha, \lambda > 0$ $\dfrac{\lambda^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\lambda x}$
$dF(x) =$	Negative Binomial $\dfrac{\Gamma(\alpha + x)}{\Gamma(\alpha)\Gamma(x+1)}\left(\dfrac{\lambda}{\lambda+1}\right)^\alpha \left(\dfrac{1}{\lambda+1}\right)^{\alpha-x}$	Pareto $\dfrac{\alpha\lambda^\alpha}{(\lambda + x)^{\alpha+1}}$
$dU(\theta\|x_1, \ldots, x_T) =$	Gamma with $\tilde{\alpha}$ and $\tilde{\lambda}$ $\dfrac{\lambda^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\lambda x}$ where $\tilde{\alpha} = \alpha + \sum_j x_j$ and $\tilde{\lambda} = \lambda + T$	Gamma with $\tilde{\alpha}$ and $\tilde{\lambda}$ $\dfrac{\lambda^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\lambda x}$ where $\tilde{\alpha} = \alpha + T$ and $\tilde{\lambda} = \lambda + \sum_j x_j$
$\mu(\theta) =$	θ	$1/\theta$
$m =$	α/λ	$\lambda/(\alpha - 1)$
$\mathcal{B}(X_1, \ldots, X_T) =$	$(\alpha + \sum_j X_j)/(\lambda + T)$	$(\lambda + \sum_j X_j)/(\alpha + T - 1)$
$z =$	$T/(T + \lambda)$	$T/(T + \alpha - 1)$

We have the following remarks on Tables 1 and 2.

1. The Poisson-gamma case yields a negative binomial distribution for X. The negative binomial distribution can be obtained from either of the following models: (i) a model without contagion but with an heterogeneous population, and (ii) a model with true contagion. Feller (1943) writes:

   > It is therefore most remarkable that Greenwood and Yule found their distribution assuming an *apparent contagion*; in their opinion this distribution *contradicts* true contagion. On the contrary, Polya and Eggenberger arrived at the *same* distribution [the negative binomial] assuming *true* contagion, while the possibility of an apparent contagion due to inhomogeneity seems not to have been noticed by them.

2. The exponential-gamma case can be generalized to a case with gamma (with parameters k and Θ) and gamma prior (with parameters α and λ). The marginal distribution of X is then a generalized Pareto distribution (Hogg and Klugman 1984). In this case, however, the Bayesian premium is no longer a credibility premium because

$$\mathcal{B}(X_1, \dots, X_T) = k \times \frac{\beta + \sum_j X_j}{\alpha + T - 1}.$$

3. In the normal/normal case, we have

$$\tilde{\sigma}_1^2 = \frac{\sigma_1^2}{T\dfrac{\sigma_1^2}{\sigma_2^2} + 1} \le \sigma_1^2,$$

   with the equality only if $\sigma_1^2 = 0$ (the case $\sigma_2^2 = \infty$ being of no interest). This inequality can be interpreted as a reduction of the uncertainty about the risk level of an insured as the amount of experience (in number of years) increases.

4. (a) In all cases, $z = T/(T + K) \to 1$ as $T \to \infty$. This is to be expected because confidence in the individual experience increases as the volume of that experience increases.

(b) In the Poisson-gamma case, where $K = \lambda$, a small value of λ means a high level of uncertainty for the value of θ (as the gamma curve will be very flat). Thus, there will be a low level of confidence in the collective premium and a high credibility factor.

(c) In the normal-normal case, K is large if σ_2^2 is large or if σ_1^2 is close to zero. If σ_2^2 is large the experience may be so volatile that one can hardly infer anything from it. When σ_1^2 is close to zero, θ is known with almost certainty. In either case, it is appropriate to charge the collective premium.

The distributions in Tables 1 and 2 are members of the so-called *unidimensional exponential family.* Jewell (1974) unified the results of Tables 1 and 2 in an elegant way. A discussion of Jewell's work, however is beyond the scope of this paper.

Goel (1982) conjectured that only combinations of unidimensional exponential family members with their *natural conjugate* priors yield linear Bayesian premiums. If Goel is correct, then the only Bayesian premiums that are exact credibility premiums are the ones found in Tables 1 and 4.

This completes the study of exact Bayesian credibility. The models of Bühlmann (1967), Bühlmann and Straub (1970), and others are based on the Bayesian approach to credibility. The basic mathematical model presented in this section remains valid. The main change, however, consists in the removal of the distributional assumptions so that the calculations are done in a nonparametric setting, one that is better suited to the practical applications of credibility theory.

6 The Bühlmann-Straub Model

6.1 The Model's Assumptions

The 1970 Bühlmann-Straub model is a generalization of the classical credibility model of Bühlmann (1969). It was introduced by Bühlmann and Straub as a means to rate reinsurance treaties. Since then, the model has been widely used in reinsurance or auto insurance, mostly in Europe. It forms the cornerstone of greatest accuracy credibility theory.

We consider a portfolio as depicted in Figure 4, where each line represents an insured. The portfolio is composed of I insureds each characterized by an unobservable random risk parameter Θ_i. The data consist of the available observations X_{it} for $t = 1,2,\dots,T_i$ and $i = 1,2,\dots,I$. The X_{it}s consist of relevant information of insured i's claims experience such as average claim amount or claim loss-ratio in year t. Note that the number of periods of experience, T_i, depends on the insured. To each X_{it} a weight w_{it} is assigned. The weights can be any valid measure of exposure such as the number of claims in one year or the premium volume. It is important that X_{it} is or behaves like a ratio so that its (conditional) variance will be inversely proportional to the weight assigned to X_{it}; see equation (24).

Figure 4
Illustration of the Portfolio in a Bühlmann-Straub Model

Insured	Risk Level	Annual Observations			Weights		
1	Θ_1	X_{11}	$\dots$	X_{1T_1}	w_{11}	$\dots$	w_{1T_1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	Θ_i	X_{i1}	$\dots$	X_{iT_i}	w_{i1}	$\dots$	w_{iT_i}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	Θ_I	X_{I1}	$\dots$	X_{IT_I}	w_{I1}	$\dots$	w_{IT_I}

The mathematical assumptions of the Bühlmann-Straub model are the following.

(BS1) The insureds' vectors $(X_{i1},\dots,X_{iT_i},\Theta_i)$, $i = 1,\dots,I$ are mutually independent;

(BS2) The risk parameters $\Theta_1,\dots,\Theta_I$ are independent and identically distributed;

(BS3) The variables X_{it} have finite variance; and

(BS4) For $i = 1,\dots,I$ and $t,u = 1,\dots,T_i$,

$$\begin{aligned} \mathrm{E}[X_{it}|\Theta_i] &= \mu(\Theta_i), && (23) \\ \mathrm{Cov}(X_{it},X_{iu}|\Theta_i) &= \delta_{tu}\frac{\sigma^2(\Theta_i)}{w_{it}} && (24) \end{aligned}$$

where δ_{tu} is the Kronecker delta, which equals one if $t = u$ and zero otherwise. Note that equation (24) states that, given the risk parameter, successive claim records of an insured are uncorrelated. Complete independence is thus not required. Claim records are nevertheless correlated unconditionally.

While assumption (BS1) represents the independence *between* the insureds, equation (24) reflects the noncorrelation *within* the insureds' claims experience across the years and the homogeneity in time. Indeed, one notes that the risk premium $\mu(\Theta_i)$ is time invariant. If the X_{it}s represent claim ratios, the claim amounts must be deflated to remove any trend in the data. If the data, nevertheless, show a trend, then a regression model like the one of Hachemeister (1975) should be used instead of the Bühlmann-Straub model.

6.2 Estimation of the Credibility Premium

Following Bühlmann (1967), the estimator of the risk premium is restricted to be a *linear* Bayesian premium. In the Bühlmann-Straub model, this linear Bayesian premium also happens to be a credibility premium. The notation used is as follows:

$$\begin{aligned}
m &= \mathrm{E}[\mu(\Theta_i)] \\
s^2 &= \mathrm{E}\left[\sigma^2(\Theta_i)\right] \\
a &= \mathrm{Var}[\mu(\Theta_i)] \\
w_{i\bullet} &= \sum_{t=1}^{T_i} w_{it} \\
w_{\bullet\bullet} &= \sum_{i=1}^{I}\sum_{t=1}^{T_i} w_{it} \\
X_{i\bullet}^{(w)} &= \sum_{t=1}^{T_i} \frac{w_{it}}{w_{i\bullet}} X_{it} \\
X_{\bullet\bullet}^{(w)} &= \sum_{i=1}^{I}\sum_{t=1}^{T_i} \frac{w_{it}}{w_{\bullet\bullet}} X_{it} \\
z_{\bullet} &= \sum_{i=1}^{I} z_i \\
X_{\bullet\bullet}^{(zw)} &= \sum_{i=1}^{I} \frac{z_i}{z_{\bullet}} \sum_{t=1}^{T_i} \frac{w_{it}}{w_{\bullet\bullet}} X_{it}.
\end{aligned}$$

The term z_i is called the credibility factor, $X_{i\bullet}^{(w)}$ is a weighted average of the claims experience of insured i. The terms m, s^2, and a are called the *structure parameters.* Notice that they are independent of i because of assumption (BS3). These structure parameters are generally unknown and must be estimated from the entire portfolio data.

The credibility premium, $\mathcal{P}_i$, is the estimator that is closest to $\mu(\Theta_i)$ or to X_{i,T_i+1} in the sense of minimizing the mean square error, i.e.,

$$\mathcal{P}_i(X_{i1},\ldots,X_{iT_i}) = \min_{\pi(\cdot)} \mathrm{E}\left[(\mu(\Theta_i) - \pi(X_{i1},\ldots,X_{iT_i})^2)\right] \tag{25}$$

or

$$\hat{X}_{i,T_i+1} = \min_{y(\cdot)} \mathrm{E}\left[(X_{i,T_i+1} - y(X_{i1},\ldots,X_{iT_i}))^2\right] \tag{26}$$

where both $\pi(\cdot)$ and $y(\cdot)$ are linear functions of the data.

The solution (see, for example, Goovaerts and Hoogstad (1987)) to equations (25) and (26) is:

$$\mathcal{P}_i(X_{i1},\ldots,X_{iT_i}) \equiv \hat{X}_{i,T_i+1} = z_i X_{i\bullet}^{(w)} + (1 - z_i)m, \tag{27}$$

where

$$z_i = \frac{w_{i\bullet}}{w_{i\bullet} + K}, \quad K = \frac{s^2}{a}. \tag{28}$$

From the definition of the credibility premium, there is no point in artificially increasing the credibility factors. Indeed, given the true values of m, s^2, and a, the factors calculated with equation (28) yield the closest estimates to insureds' risk premiums.

The structure parameter s^2 is a global measure of the stability of the portfolio's claim experience; s^2 is sometimes called the *homogeneity within the insureds.* The lower the value of s^2 the more stable the portfolio's claim experience is, and, as in limited fluctuation credibility, the larger the credibility factor.

The structure parameter a is a measure of the variation of the various individual risk premiums and is sometimes referred to as *the homogeneity between the insureds.* In other words, a is an indicator of the heterogeneity of the portfolio's experience. The greater the heterogeneity of a portfolio, the more important is the weight given to individual experience. Hence, as a increases, the z_is increase also.

For further discussion of the interpretations of s^2 and a see the target and shooters example of Philbrick (1981).

6.3 Estimation of the Structure Parameters

The structure parameters m, s^2, and a are functionals of the unobservable random variable Θ and are unknown in practice. Hence they must be estimated from the entire portfolio data.

6.3.1 Estimation of m

The obvious unbiased estimator of the collective premium m is $\hat{m}_1$, the average of the individual premiums weighted by their natural weights w_{it}, i.e.,

$$\hat{m}_1 = X_{\bullet\bullet}^{(w)}. \tag{29}$$

This is the estimator used in a classical statistical model (i.e., $a = 0$).

De Vylder (1978) shows that, in credibility theory, the estimator with minimum variance in the class of all unbiased linear estimators is not $\hat{m}_1$. Rather it is $\hat{m}_2$, the average of the individual premiums weighted by the credibility factors:

$$\hat{m}_2 = X_{\bullet\bullet}^{(zw)}. \tag{30}$$

The estimator $\hat{m}_2$ is called a *pseudo-estimator* because it is a function of the unknown parameters s^2 and a. When the credibility factors are known, $\hat{m}_2$ is unbiased. On the other hand, it is not known if $\hat{m}_2$ is unbiased when the credibility factors are unknown. Various practical tests made by the author and others[7] show that the estimator $\hat{m}_2$ is more precise and reliable than $\hat{m}_1$, provided the parameters s^2 and a are suitably estimated. Thus $\hat{m}_2$ is used as our estimator of m.

One more point in favor of $\hat{m}_2$ is that it ensures that enough premiums are collected to cover the expected losses, i.e., the equivalence principle is verified. Using equation (27), mathematically, this means that

[7] See, for example, Goovaerts, Kaas, van Heerwaarden, and Bauwelinckx (1990) for more on these tests.

$$\begin{aligned}\sum_{i=1}^{I} w_{i\bullet}\mathcal{P}_i(X_{i1},\dots,X_{iT_i}) &\equiv \sum_{i=1}^{I} w_{i\bullet}\hat{X}_{i,T_i+1}\\ &= \sum_{i=1}^{I} w_{i\bullet}[z_i X_{i\bullet}^{(w)} + (1-z_i)\hat{m}_2]\\ &= \sum_{i=1}^{I} w_{i\bullet}X_{i\bullet}^{(w)}.\end{aligned}$$

6.3.2 Estimation of s^2

The estimation of the structure parameter s^2 is fairly straightforward. A good unbiased estimator of s^2 is

$$\hat{s}^2 = \frac{1}{N-I}\sum_{i=1}^{I}\sum_{t=1}^{T_i} w_{it}(X_{it} - X_{i\bullet}^{(w)})^2, \quad \text{where } N = \sum_{i=1}^{I} T_i. \tag{31}$$

Dubey and Gisler (1981), in an excellent paper on parameter estimation in credibility theory, consider some variants of $\hat{s}^2$ that prove inferior. De Vylder and Goovaerts (1992) show that the estimator $\hat{s}^2$ in equation (31) is optimal (has minimum variance) in a wide class of estimators, if the conditional random variables $X_{it}|\Theta_i$ have a coefficient of excess equal to zero. The coefficient of excess (also called the kurtosis) γ_2 of a random variable Y is defined as

$$\gamma_2(Y) = \frac{\mathrm{E}[(Y-\mathrm{E}[Y])^4]}{\mathrm{E}[(Y-\mathrm{E}[Y])^2]^2} - 3.$$

A normal random variable has zero-excess, i.e., $\gamma_2 = 0$.

6.3.3 Estimation of a

There are two main unbiased estimators for parameter a. The first one is derived from the ANOVA (analysis of variance) between sample variance; this estimator is denoted by $\hat{a}$:

$$\hat{a} = \frac{w_{\bullet\bullet}}{w_{\bullet\bullet}^2 - \sum_i w_{i\bullet}^2}\left(\sum_{i=1}^{I} w_{i\bullet}(X_{i\bullet}^{(w)} - X_{\bullet\bullet}^{(w)})^2 - (I-1)s^2\right). \tag{32}$$

This estimator can be negative, which is a drawback. If one uses $\hat{a}' = \max(\hat{a}, 0)$ instead, then the estimator $\hat{a}'$ will be biased.

The second estimator, generally known as the Bichsel-Straub estimator, is denoted $\tilde{a}$:

$$\tilde{a} = \frac{1}{I-1} \sum_{i=1}^{I} z_i (X_{i\bullet}^{(w)} - X_{zw})^2. \tag{33}$$

This estimator is always positive. Unfortunately, the right side of equation (33) contains parameter a (via the credibility factors), making $\tilde{a}$ a pseudo-estimator. Thus $\tilde{a}$ must be calculated iteratively.

Dubey and Gisler (1981) demonstrated that when $\tilde{a}$ is calculated iteratively using a fixed-point iterative scheme then $\tilde{a}$ converges to a strictly positive value whenever $\hat{a} > 0$. Otherwise, $\tilde{a}$ necessarily converges to 0.

If the variables $X_{i\bullet}^{(w)}$ have zero coefficient of excess, De Vylder and Goovaerts (1992) show that $\tilde{a}$ has minimum variance for a wide class of estimators. For the estimation of structure parameters, especially under zero-excess assumptions, one can also refer to De Vylder (1996).[8]

The following algorithm illustrates the complete estimating procedure: let $\tilde{a}_k$ denote the estimate of $\tilde{a}$ during the kth iteration.

1. Calculate $\hat{s}^2$ with equation (31);
2. Calculate $\hat{a}$ with equation (32). If the resulting value is negative, put $\hat{a} = 0$;
3. If $\hat{a} > 0$ then
   (a) Set $\tilde{a}_0 = \hat{a}$;
   (b) Obtain a new value $\tilde{a}_{n+1}$ with equation (33) using $\tilde{a}_n$, $n = 0, 1, \ldots$;
   (c) Repeat Step 3b until $|\tilde{a}_{n+1} - \tilde{a}_n|$ or $|\tilde{a}_{n+1} - \tilde{a}_n| / \tilde{a}_n$ is sufficiently small.

   Else, put $\tilde{a} = 0$
4. Calculate the credibility factors with equation (28) using $\hat{s}^2$ and $\hat{a}$ or $\tilde{a}$;

[8]The author has observed in practice, and De Vylder and Goovaerts (1991) confirm it by theoretical arguments, that $\hat{a}$ appears to be more accurate when a *small* value of a is expected and vice-versa.

5. Calculate the credibility premium $\mathcal{P}_i(X_{i1}, \dots, X_{iT_i})$ with equation (30).

This procedure will now be illustrated by a numerical example.

6.4 A Numerical Example

The data used in this numerical example are obtained via the simulation of an insurance portfolio according to the Bühlmann-Straub model assumptions. The portfolio consists of 20 individuals ($I = 20$) and the simulation period is 5 years ($T_i = 5$).

- First, weights are drawn from a uniform distribution on (c, d) where $0 \le c < d$;
- Next, the risk levels, Θ_i, are simulated from a gamma distribution with parameters α and λ;
- The total number of claims for insured i in year t, N_{it}, is obtained from a Poisson distribution with parameter $w_{it}\Theta_i$;
- Each of these N_{it} claims is simulated from a gamma distribution with parameters γ and β, i.e., the kth claim, Y_{itk} for $k = 1, 2, \dots, N_{it}$ has a gamma distribution with parameters γ and β;
- The total claim amount for insured i in year t, $S_{it}|\Theta_i$ is then equal to the sum of N_{it} claims, i.e.,

$$S_{it}|\Theta_i = \sum_{k=1}^{N_{it}} Y_{itk}; \quad \text{and}$$

- Finally, the ratios are calculated by dividing the total claim amounts by the weights, i.e., $X_{it} = S_{it}/w_{it}$.

Note that $S_{it}|\Theta_i$ follows a compound Poisson distribution.

The structure parameters are thus given by

$$\begin{aligned}
m &= \mathrm{E}[Y]\,\mathrm{E}[\Theta_i] = \frac{\gamma\alpha}{\beta\lambda} \\
s^2 &= \mathrm{E}\left[Y^2\right]\mathrm{E}[\Theta_i] = \frac{\gamma(\gamma+1)\alpha}{\beta^2\lambda} \\
a &= \mathrm{E}[Y]^2\,\mathrm{Var}[\Theta_i] = \frac{\gamma^2\alpha}{\beta^2\lambda^2}.
\end{aligned}$$

The parameter values used in this example are: $c = 1,000$, $d = 200,000$, $\gamma = 7$, $\beta = 0.002$, $\alpha = 5$, and $\lambda = 10,000$. This results in theoretical values for the structure parameters of $m = 1.75$, $s^2 = 7,000$, and $a = 0.6125$.

Tables 3 and 4 display the results of the simulation for the first five years. Using the estimators described in Section 6.2 yields the following estimates: $\hat{s}^2 = 6,971$, $\hat{a} = 0.6426$, $\tilde{a} = 0.6136$, $K = 10,848$, and $\hat{m}_2 = 1.7446$ (with $\hat{a}$), $K = 11,361$, and $\hat{m}_2 = 1.7447$ (with $\tilde{a}$). The credibility premiums in Table 3 are calculated using $\tilde{a}$.

For example, insured #1 has a total amount of claims of 1,011,179 and total exposure of 624,100 in the first five years. This insured's credibility factor is thus $624,100/(624,100 + 11,361) = 0.9821$ and yields a credibility premium for the sixth year of 1.6224 $(0.9821(1,6202) + (1 - 0.9821)(1.7447) = 1.6224)$.

By comparing the actual sixth year ratios with the credibility premiums, one can measure the precision of these premiums. The mean squared error for the credibility premiums is equal to 0.128. Using instead the individual means, $X_{i\bullet}^{(w)}$, the mean squared error increases (as expected) to 0.132.

6.5 Some Practical Issues

One problem that may be encountered in practice when using the Bühlmann-Straub model is possible annual variations in $K = s^2/a$. In theory, these variations may well be appropriate if the structure of the portfolio changes. If the ratemaking procedure is transparent in some way, however, a company may be reluctant to significantly change the credibility factor of an insured from one year to the next. To deal with this problem, Sundt (1992) proposes an interesting solution combining greatest accuracy and limited fluctuation credibility to actually decrease the rate of the convergence of the credibility premium to the true risk premium. The reader is encouraged to read Sundt's paper for more details.

Another potential problem is that of outliers. Like all variance estimators, $\hat{s}^2$, $\hat{a}$, and $\hat{a}$ are affected by extreme values called *outliers*. For example, one outlier, even if only lightly weighted, may have a significant effect on the calculation of $\hat{s}^2$ and, in addition, may cause $\hat{a}$ to be negative.

Table 3
Observed Ratios X_{it} for 20 Insureds

	First 5 Years Ratios					6th Year Ratios	
(1)	1	2	3	4	5	Actual	CREDIB
1	1.265	1.749	1.501	2.075	1.574	1.650	1.6224
2	0.966	0.497	1.010	1.016	0.865	1.172	0.9482
3	1.049	1.081	1.591	1.118	0.916	0.857	1.0943
4	3.729	2.185	2.833	3.308	2.980	2.279	3.0013
5	0.958	2.094	1.883	1.580	2.056	2.324	1.8807
6	2.886	3.166	3.021	3.441	2.716	2.979	3.0106
7	2.182	1.693	1.809	1.904	1.859	1.408	1.9331
8	1.674	1.606	1.386	1.558	1.664	1.232	1.5779
9	1.143	1.024	1.071	1.237	1.330	1.268	1.1718
10	1.829	2.083	1.926	2.737	2.434	2.944	2.2824
11	0.703	1.367	1.189	1.509	1.058	0.721	1.0561
12	1.733	1.582	1.505	1.627	0.983	1.681	1.4575
13	1.664	1.714	1.573	1.639	1.752	1.608	1.6574
14	0.859	0.453	0.805	0.605	0.706	0.790	0.7482
15	2.111	2.697	2.312	2.985	2.880	2.145	2.6630
16	1.320	1.408	1.189	1.437	1.145	1.334	1.3113
17	3.750	2.756	3.530	3.502	3.083	2.945	3.4069
18	0.594	0.721	1.208	0.962	0.191	1.021	0.9035
19	2.058	2.048	2.251	1.579	1.850	2.833	2.0053
20	1.181	1.485	0.620	1.474	0.860	0.916	1.1623

Notes: Column (1) lists the 20 insureds, $I = 1, 2, \ldots, 20$; CREDIB = Credibility premiums calculated from the previous five years of observed ratios.

Table 4
Weights w_{it} Used in the Numerical Example

	Observed weights					
(1)	1	2	3	4	5	6
1	58700	169200	177900	60600	157700	196700
2	163500	41800	156000	152600	157500	92100
3	127200	102700	8600	177500	49100	30900
4	64000	39600	106900	69700	157500	85600
5	11300	76600	95600	127000	191800	101600
6	126100	16800	177500	133700	108300	39300
7	168400	76500	102500	51800	97200	72300
8	60600	53900	124500	126300	199300	79500
9	168600	131100	15400	84300	87000	74800
10	57500	177300	125300	182200	193600	127900
11	170600	124900	11600	26300	73100	9900
12	40200	49400	74400	77700	78400	75200
13	149900	144600	143600	65800	33400	97100
14	139300	27800	152800	146200	148400	135600
15	67800	64600	126700	190200	133100	12500
16	150700	100600	140100	80700	54100	166700
17	145600	7900	170000	182400	198300	72100
18	80600	59100	88600	120200	11000	47700
19	148200	165400	153800	48400	187100	33300
20	138100	78100	39100	102000	148900	88900

Notes: Column (1) lists the 20 insureds, $I = 1, 2, \ldots, 20$.

If the insured under study is simply removed from the data set, the estimators typically will immediately revert to more standard values for the portfolio. Consequently, if a portfolio's claim distribution or ratio distribution is highly skewed to the right, it is generally preferable to modify the data in some way to reduce the effect of outliers on the estimators.

It is not the purpose of this paper to discuss these considerations in any detail. The interested reader can review the semilinear model of De Vylder (1976b), which uses a special transformation of the observations (X_{it}), the optimal trimming procedure of Gisler (1980), and the robust estimators of Gisler and Reinhard (1993).

7 Hierarchical Credibility

In the credibility premium for insured i (defined in (27)), only the data from insured i are used in favor of the (assumed known) structure parameters m, s^2, and a. The entire portfolio data are used only if the structure parameters are unknown. The reason for using only data from insured i is because of the assumed independence between the various insureds' risk levels. This non-utilization of useful collateral data that may contain information on the risk level of insured i disturbed some Bayesian theoreticians when the Bühlmann (1967) credibility model was introduced. In response, Jewell (1975) developed a solution: the hierarchical credibility model.

Today, hierarchical credibility theory is seen as an efficient way to apply credibility theory to very large portfolios. It is a generalization of almost any single level credibility model such as the Bühlmann-Straub (1970) model, the Hachemeister (1975) regression model, and the De Vylder (1976a) semi-linear model, to name only a few. It is worth mentioning that extensions of the hierarchical model to the regression case are due to Sundt (1979, 1980) and that the fully general scheme is due to Norberg (1986).

As the number of insureds in a portfolio increases, the portfolio may become too heterogeneous to be successfully rated. The fundamental idea of hierarchical credibility theory is to divide large portfolios into smaller more homogeneous subportfolios under some general criteria. For example, in territorial automobile rating, the portfolio can be subdivided according to state or province (the upper level), then by county (second level), then by size of city (third level), with the driver at the lowest level. The end result is a tree-like (hierarchical) classification structure similar to the one displayed in Figure 5, which depicts a four-level hierarchical classification. In Figure 5 the upper most level is the entire portfolio, the next level depicts the *sectors*, one step below is the *classes*. The final and lowest level contains the insureds.[9]

7.1 General Presentation of the Model

For convenience and without loss of generality, the discussion of the hierarchical model is based on the four-level hierarchical model displayed in Figure 5. The presentation follows Bühlmann and Jewell (1987) so most of their terminology is used. Again, for convenience, weights are specifically added to our formulae. To avoid the proliferation of subscripts, the risk parameter of each level is represented by a different letter.

The various levels of the model are the following.

- **Level 4**: This is the entire portfolio.
- **Level 3**: The portfolio is divided into sectors. The risk level of the pth sector is unobservable and is denoted by ψ_p for $p = 1, \ldots, P$. The ψ_ps are assumed to be realizations of the random variable Ψ.
- **Level 2**: Each sector is further divided into classes. The kth class of the pth sector is characterized by an unobservable risk level $\phi_k^{(p)}$, $k = 1, \ldots, K_p$ and $p = 1, \ldots, P$. For a given sector p, the $\phi_k^{(p)}$s are assumed to be realizations of the random variable

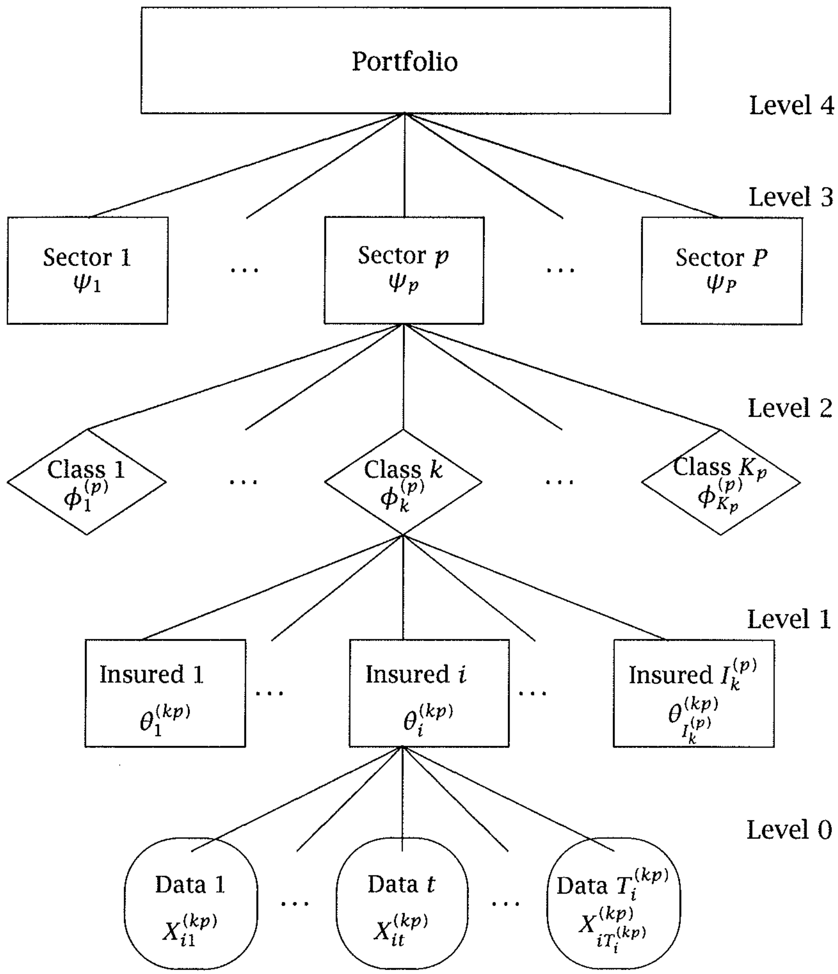
$$\Phi^{(p)} = \Phi | \Psi = \psi_p.$$

- **Level 1**: Each class consists of homogeneous insureds. Each insured has an unobservable risk level $\theta_i^{(kp)}$, for $i = 1, \ldots, I_k^{(p)}$, $k = 1, \ldots, K_p$ and $p = 1, \ldots, P$. For a given sector p and class k, the $\theta_i^{(kp)}$s are assumed to be realizations of the random variable

$$\Theta_i^{(kp)} = \{\Theta | \Phi = \phi_k^{(p)} \bigcap \Psi = \psi_p\}.$$

[9]The structure depicted in Figure 5 is similar to the classification structure of the workers compensation board in Quebec, Canada.

Figure 5
Graphical Representation of a Four-Level Hierarchical Model



- **Level 0**: This consists of the raw data. The dataset for insured i in class k and sector p level of the portfolio data is denoted by $\mathcal{D}_i^{(kp)}$ and the corresponding set of weights is $\mathcal{W}_i^{(kp)}$ where

$$\begin{aligned} \mathcal{D}_i^{(kp)} &= (X_{i1}^{(kp)}, \ldots, X_{iT_i^{(kp)}}^{(kp)}) \\ \mathcal{W}_i^{(kp)} &= (w_{i1}^{(kp)}, \ldots, w_{iT_i^{(kp)}}^{(kp)}) \end{aligned}$$

for $i = 1, \ldots, I_k^{(p)}$, $k = 1, \ldots, K_p$ and $p = 1, \ldots, P$.

The mathematical assumptions of the hierarchical model are the following:

(H1) The variables Ψ_p are i.i.d. with cumulative distribution function (cdf) $H_3(\cdot)$;

(H2) For given p, the variables $\Phi_k^{(p)}$ are conditionally i.i.d. with cdf $H_2(\cdot|\psi_p)$;

(H3) For given p and k, the variables $\Theta_i^{(kp)}$ are conditionally i.i.d. with cdf $H_1(\cdot|\phi_k^{(p)})$;

(H4) For a given sector p, class k, and $\theta_i^{(kp)}$, the $X_{i1}^{(kp)}, X_{i2}^{(kp)}, \ldots, X_{iT_i^{(kp)}}^{(kp)}$ are i.i.d. and have finite variance;

(H5) For all $i = 1, \ldots, I_k^{(p)}$ and $t, u = 1, \ldots, T_i^{(kp)}$,

$$\begin{aligned} \mathrm{E}\left[X_{it}^{(kp)}|\Theta_i^{(kp)}\right] &= \mu(\Theta_i^{(kp)}), \\ \mathrm{Cov}(X_{it}^{(kp)}, X_{iu}^{(kp)}|\Theta_i^{(kp)}) &= \delta_{tu}\frac{\sigma^2(\Theta_i^{(kp)})}{w_{it}^{(kp)}}. \end{aligned}$$

7.2 Credibility Estimates

Let us define the risk premium and the various structure parameters at each level:

- **Level 0:** Bühlmann and Jewell define the *linearly sufficient statistic* for this level as

$$B_{kp}(\Theta_i^{(kp)}) = X_{i\bullet}^{(wkp)} \tag{34}$$

where

$$X_{i\bullet}^{(wkp)} = \sum_{t=1}^{T_i^{(kp)}} \frac{w_{it}^{(kp)}}{w_{i\bullet}^{(kp)}} X_{it}^{(kp)} \tag{35}$$

and

$$w_{i\bullet}^{(kp)} = \sum_{t=1}^{T_i^{(kp)}} w_{it}^{(kp)} \tag{36}$$

- **Level 1:**

$$\mu(\Theta_i^{(kp)}) \quad \text{and} \quad \sigma^2(\Theta_i^{(kp)})$$

- **Level 2:**

$$\begin{aligned}
M_{kp}(\Phi_k^{(p)}) &= \mathrm{E}\left[\mu(\Theta_i^{(kp)})|\Phi_k^{(p)}\right], \\
F_{kp}(\Phi_k^{(p)}) &= \mathrm{E}\left[\sigma^2(\Theta_i^{(kp)})|\Phi_k^{(p)}\right],
\end{aligned}$$

and

$$G_{kp}(\Phi_k^{(p)}) = \mathrm{Var}\left[\mu(\Theta_i^{(kp)})|\Phi_k^{(p)}\right].$$

- **Level 3:**

$$\begin{aligned}
M_p(\Psi_p) &= \mathrm{E}\left[M_{kp}(\Phi_k^{(p)})|\Psi_p\right] \\
F_p(\Psi_p) &= \mathrm{E}\left[F_{kp}(\Phi_k^{(p)})|\Psi_p\right], \\
G_p(\Psi_p) &= \mathrm{E}\left[G_{kp}(\Phi_k^{(p)})|\Psi_p\right],
\end{aligned}$$

and

$$H_p(\Psi_p) = \mathrm{Var}\left[H_{kp}(\Phi_k^{(p)})|\Psi_p\right].$$

- **Level 4:**

$$
\begin{aligned}
M &= \mathrm{E}\Big[M_p(\Psi_p)\Big], \\
F &= \mathrm{E}\Big[F_p(\Psi_p)\Big], \\
G &= \mathrm{E}\Big[G_p(\Psi_p)\Big], \\
H &= \mathrm{E}\Big[H_p(\Psi_p)\Big],
\end{aligned}
$$

and

$$I = \mathrm{Var}\Big[M_p(\Psi_p)\Big],$$

which are constants.

Like the standard Bühlmann-Straub model, the goal here is to find the credibility premium closest (in the mean squared error sense) to $\mu(\Theta_i^{(kp)})$. This requires, however, the estimation of $M_{kp}(\Phi_k^{(p)})$, $M_p(\Psi_p)$, and M, which are the class, sector, and portfolio risk premiums, respectively. The credibility premium in the hierarchical model, $\mathcal{P}_i^{(kp)}$, is determined as follows:

$$
\begin{aligned}
\mathcal{P}_i^{(kp)} &= z_i^{(kp)} B_i^{(kp)} + (1 - z_i^{(kp)})\hat{M}_{kp}, && (37) \\
\hat{M}_{kp} &= z_k^{(p)} B_k^{(p)} + (1 - z_k^{(p)})\hat{M}_k, && (38) \\
\hat{M}_k &= z_p B_p + (1 - z_p)\hat{M}, && (39)
\end{aligned}
$$

where $B_i^{(kp)}$ is defined in equation (34)

$$
\begin{aligned}
B_k^{(p)} &= \sum_{i=1}^{I_k^{(p)}} \frac{z_i^{(kp)}}{z_\bullet^{(kp)}} B_i^{(kp)}, && (40) \\
B_p &= \sum_{k=1}^{K_p} \frac{z_k^{(p)}}{z_\bullet^{(p)}} B_k^{(p)} && (41)
\end{aligned}
$$

and

$$
\begin{aligned}
z_i^{(kp)} &= \frac{V_i^{(kp)}}{V_i^{(kp)} + F/G}, \quad V_i^{(kp)} = w_{i\bullet}^{(kp)}, && (42) \\
z_k^{(p)} &= \frac{V_k^{(p)}}{V_k^{(p)} + G/H}, \quad V_k^{(p)} = z_\bullet^{(p)} = \sum_{i=1}^{I_k^{(p)}} z_i^{(kp)}, && (43) \\
z_p &= \frac{V_p}{V_p + H/I}, \quad V_p = z_\bullet^{(p)} = \sum_{k=1}^{K_p} z_k^{(p)}. && (44)
\end{aligned}
$$

In our notation, the function B is used to represent the linearly sufficient statistic of a level while the function V represents its total volume. A closer look at the above formulae shows that, except for level 1, where the natural weights are used, the following observations can be made:

- The total volume of a level is the sum of the credibility factors of the previous level, and
- The linearly sufficient statistic of a level is the credibility-weighted average of the linearly sufficient statistics of the previous level.

The credibility premium in a hierarchical model, $\mathcal{P}_i^{(kp)}$, depends on several unknown structure parameters: the portfolio mean M and the variance parameters F, G, H, and I. The pseudo-estimators below are all unbiased; see Goovaerts et al., 1990:

$$\hat{M} = \sum_{p=1}^{P} \frac{z_p}{z_\bullet} B_p \tag{45}$$

$$= \sum_{p=1}^{P} \sum_{k=1}^{K_p} \sum_{i=1}^{I_k^{(p)}} \sum_{t=1}^{T_i^{(kp)}} \frac{z_p^{(3)}}{z_\bullet^{(3)}} \frac{z_k^{(2)}}{z_\bullet^{(2)}} \frac{z_i^{(1)}}{z_\bullet^{(1)}} \frac{w_{it}}{w_{i\bullet}} X_{it}, \tag{46}$$

$$z_\bullet = \sum_{p=1}^{P} z_p \tag{47}$$

$$\hat{F} = \frac{1}{d_1} \sum_{p,k,i,t} w_{it}^{(kp)} (X_{it}^{(kp)} - B_i^{(kp)})^2, \tag{48}$$

$$\hat{G} = \frac{1}{d_2} \sum_{p,k,i} z_i^{(kp)} (B_i^{(kp)} - B_k^{(p)})^2, \tag{49}$$

$$\hat{H} = \frac{1}{d_3} \sum_{p,k} z_k^{(p)} (B_k^{(p)} - B_p)^2, \tag{50}$$

$$\hat{I} = \frac{1}{d_4} \sum_{p} z_p^{(3)} (B_p - \hat{M})^2, \tag{51}$$

where the denominators d_1 through d_4 are equal to the total number of terms in the corresponding sum minus the number of terms B, i.e., they are similar to the numbers of degrees of freedom.

7.3 Interpretation of the Results

One important point to note about the hierarchical model is that it is completely different from the Bühlmann-Straub model. If, for example, the Bühlmann-Straub model was applied to each class separately, the credibility constant K would vary by class, i.e., we would have

$$K_{kp} = s_{kp}^2 / a_{kp}$$

where s_{kp}^2 and a_{kp} are defined with respect to the kth class in the pth sector. In a hierarchical model, on the other hand, the credibility constants F/G and G/H in equation (42) and equation (43) are the same for, respectively, every class and sector of the portfolio. Any two insureds (or classes) with the same total weight consequently must have identical credibility factors.

Though the Bühlmann-Straub model (applied on a class-by-class basis) and the hierarchical model are theoretically valid approaches, they are not equivalent. One must make an enlightened choice of one approach over the other. The Bühlmann-Straub model assumes that all classes, regardless of sector, are mutually independent.

The hierarchical model, on the other hand, purposely creates a dependency between insureds of different classes and sectors. By choosing this model, one therefore considers that sectors, classes, and insureds are not completely independent; that is, the data of any one insured bears some (indirect) ratemaking information about another insured in a different class or sector. This is just the use of collateral data for which the model was created.

In addition, the assumption that the risk levels are conditionally i.i.d. implies that, a priori, nothing is known about the relative risk levels of the sectors and classes. That is, if one knows *with certainty* that a given sector constitutes a worse risk than the others, then that sector must be excluded from the hierarchical model and be rated separately. It is the ignorance of these risk levels that leads to portfolio-wide credibility constants.

Now that this important distinction is made, we can look at two other properties of the hierarchical model. First, the model described here is not equivalent to the more intuitive approach of summing all claims and weights of a class to calculate its credibility premium. The total weight of a class (or sector) is given by the sum of its credibility factors, rather than the sum of its natural weights.

Second, when constructing hierarchical classifications, homogeneous classes of individuals are evidently desirable. The situation is reversed when it comes to sectors, because homogeneous sectors reflect an inadequate classification structure creating classes that are too similar. Homogeneous sectors can also signal an insufficient number of classes. The same arguments hold for the heterogeneity between sectors. Note that homogeneity *between* the components of a particular level is the homogeneity *within* the level above.

The interested reader may find more information on the hierarchical model and other one-level models in Goovaerts, Kaas, van Heerwaarden and Bauwelinckx (1990), Goovaerts and Hoogstad (1987), and De Vylder (1996).

8 Crossed Classification Credibility

While the hierarchical model successfully generalizes most one-level models by allowing complex tree-like classification structures, it is of little help if there is interaction between the various risk factors. For example, one can think of an automobile insurance portfolio that would be hierarchically classified first by gender of the driver and then by age. Gender and age represent here two qualitative risk factors. Without a doubt, some young women share risk characteristics with young men, who themselves share driving characteristics with older men. The hierarchical classification is therefore inappropriate in such a case. The appropriate model in such a situation is the so-called crossed classification credibility model of Dannenburg (1995), which generalizes the hierarchical model, although with some restrictions to be mentioned later. A prerequisite to the study of the crossed classification credibility model is the presentation of variance components models.

8.1 Variance Components Models

Variance components models are derived from the statistical theory of linear models; see, for example, Searle (1971) for linear models in general and Searle, Casella, and McCulloch (1992) for variance components models in particular. Dannenburg (1995) introduced variance components models in credibility as a means of generalizing the results of hierarchical credibility theory. Dannenburg, Kaas and Goovaerts (1996) then used it to present many of the classic credibility models in new and original ways—for credibility theory at least. In the variance components approach to credibility, the insured's claim ratio is decomposed into a sum of uncorrelated random variables. Each of these new random variables represents the contribution of a risk factor or the contribution of an interaction between risk factors to the total variance of the insured's claim ratio random variable. This total variance is thus broken up into so-called *variance components.*

For example, in the Bühlmann-Straub model, X_{it} can be written as

$$X_{it} = m + \Xi_i^{(1)} + \Xi_{it}^{(12)},$$

where $\Xi_i^{(1)}$ represents the variability between the insureds and $\Xi_{it}^{(12)}$ represents the variability in the insured's claims over time. The means and variances are:

$$\begin{aligned} \mathrm{E}\left[\Xi_i^{(1)}\right] &= 0, \\ \mathrm{E}\left[\Xi_{it}^{(12)}\right] &= 0, \\ \mathrm{Var}\left[\Xi_i^{(1)}\right] &= a, \\ \mathrm{Var}\left[\Xi_{it}^{(12)}\right] &= \frac{s^2}{w_{it}}. \end{aligned}$$

The random variables $\Xi_i^{(1)}$ and $\Xi_{it}^{(12)}$ are assumed to be mutually independent. (This assumption of independence can be reduced to an assumption of no correlation if only the first two moments are dealt with in the sequel.) The ratio of an insured is therefore now seen as a random variation around the collective mean m.

Another example is the hierarchical model where each level corresponds to a risk factor: sector is component 1, class is component 2, insured is component 3, and time is component 4. At time t the claim ratio of insured i in class k of sector p can be written as:

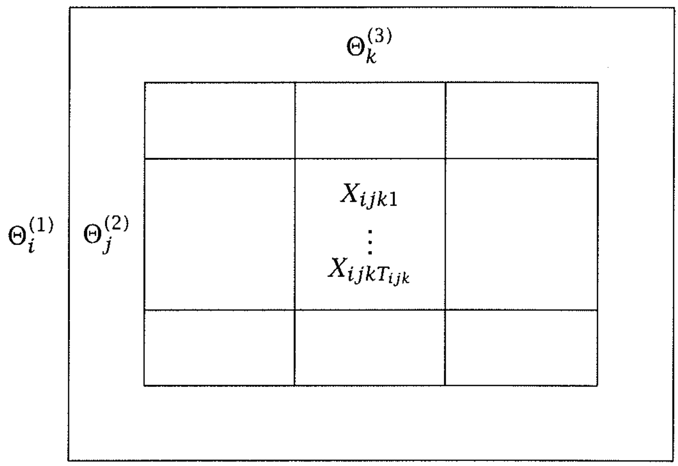
$$X_{it}^{(kp)} = m + \Xi_p^{(1)} + \Xi_{pk}^{(12)} + \Xi_{pki}^{(123)} + \Xi_{pkit}^{(1234)},$$

where a superscript of the form (i) denotes a component, (ij) denotes an interaction between components i and j, (ijk) denotes an interaction between components i, j and k, etc. The variable t is always considered as a last additional risk factor appearing only in interaction with all the other risk factors.

8.2 The Crossed Classification Credibility Model

In order to avoid an unduly cumbersome notation, a three-way (or three-factor) crossed classification credibility model is presented. This model is displayed in Figure 6.

Figure 6
A Three-Way Crossed Classification Credibility Model



In the three-way model, each risk in the portfolio is affected by three qualitative risk factors (*levels* in the hierarchical model, or *ways*) and all possible interactions between these risk factors. Though time is considered as the fourth risk factor, time is not included in the interaction terms except when all of the other factors are included.

Category $i = 1, \ldots, I$ of the first risk factor is characterized by the random variable $\Theta_i^{(1)}$, category $j = 1, \ldots, J$ of the second risk factor is characterized by the random variable $\Theta_j^{(2)}$, and category $k = 1, \ldots, K$ of the third risk factor is characterized by the random variable $\Theta_k^{(3)}$. The time subscript t goes from 1 to T_{ijk} to allow for variation in the amount of experience from one insured to another.

The assumptions of the model are as follows:

(CCC1) The elements of each of the following random vectors are i.i.d. within their respective vectors:

1. $\underline{\Theta}^{(1)} = (\Theta_1^{(1)}, \ldots, \Theta_I^{(1)})$
2. $\underline{\Theta}^{(2)} = (\Theta_1^{(2)}, \ldots, \Theta_J^{(2)})$
3. $\underline{\Theta}^{(3)} = (\Theta_1^{(3)}, \ldots, \Theta_K^{(3)})$.

(CCC2) The $(I + J + K)$ elements of the three random vectors $\underline{\Theta}^{(1)}$, $\underline{\Theta}^{(2)}$, and $\underline{\Theta}^{(3)}$ are mutually independent across vectors;

(CCC3) The conditional means can be expressed as:

$$\begin{aligned} \mathrm{E}\left[X_{ijkt} | \Theta_i^{(1)}, \Theta_j^{(2)}, \Theta_k^{(3)}\right] &= \mu_{123}(\Theta_i^{(1)}, \Theta_j^{(2)}, \Theta_k^{(3)}), \\ \mathrm{E}\left[X_{ijkt} | \Theta_i^{(1)}, \Theta_j^{(2)}\right] &= \mu_{12}(\Theta_i^{(1)}, \Theta_j^{(2)}), \\ \mathrm{E}\left[X_{ijkt} | \Theta_i^{(1)}\right] &= \mu_1(\Theta_i^{(1)}), \end{aligned}$$

and so on;

(CCC4) $\mathrm{E}\left[\mathrm{Cov}(X_{ijkt}, X_{pqru} | \underline{\Theta}^{(1)}, \underline{\Theta}^{(2)}, \underline{\Theta}^{(3)})\right] = \delta_{ijkt,pqru}\, s^2 / w_{ijkt}$.

Here, $\underline{\Theta}^{(1)} = (\Theta_1^{(1)}, \ldots, \Theta_I^{(1)})$ and $\delta_{ijkt,pqru}$ is still the Kronecker symbol equal to 1 if $i = p$, $j = q$, $k = r$, $t = u$, and zero otherwise.

In the most basic form of the crossed classification credibility model, X_{ijkt} is written as

$$\begin{aligned} X_{ijkt} = \; & m \\ & + \Xi_i^{(1)} + \Xi_j^{(2)} + \Xi_k^{(3)} + \Xi_{ij}^{(12)} \\ & + \Xi_{ik}^{(13)} + \Xi_{jk}^{(23)} + \Xi_{ijk}^{(123)} + \Xi_{ijkt}^{(1234)}. \end{aligned} \tag{52}$$

Given the superscripts $q, l_1, \ldots, l_q \in \{1, 2, 3\}$ and the appropriate subscripts, the random variable $\Xi^{(l_1 \ldots l_q)}$ has mean zero and variance $b^{(l_1 \ldots l_q)}$, i.e.,

$$\mathrm{E}\left[\Xi^{(l_1 \ldots l_q)}\right] = 0 \tag{53}$$

$$\mathrm{Var}\left[\Xi^{(l_1 \ldots l_q)}\right] = b^{(l_1 \ldots l_q)}, \tag{54}$$

moreover

$$\begin{aligned} \mathrm{E}\left[\Xi_{ijkt}^{(1234)}\right] &= 0 & (55)\\ \mathrm{Var}\left[\Xi_{ijkt}^{(1234)}\right] &= \frac{s^2}{w_{ijkt}}. & (56)\end{aligned}$$

The structure parameters of the model are thus the collective mean m and the variance components s^2, $b^{(1)}$, $b^{(2)}$, $b^{(3)}$, $b^{(12)}$, $b^{(13)}$, $b^{(23)}$, and $b^{(123)}$.

The credibility premium in the crossed classification credibility model is

$$\begin{aligned} \mathcal{P}_i &= \hat{X}_{ijk,T_{ijk}+1}\\ &= m + (1 - z_{ijk}^{(123)})\left(\hat{\Xi}_i^{(1)} + \hat{\Xi}_j^{(2)} + \hat{\Xi}_k^{(3)} + \hat{\Xi}_{ij}^{(12)} + \hat{\Xi}_{ik}^{(13)} + \hat{\Xi}_{jk}^{(23)}\right)\\ &\quad + z_{ijk}^{(123)}(X_{ijk}^{(w)} - m) \qquad (57)\end{aligned}$$

where the terms with the "hat" symbol (ˆ) are credibility estimators. The terms in equation (57) are defined below:

$$X_{ijk\bullet}^{(w)} = \sum_{t=1}^{T_{ijk}} \frac{w_{ijkt}}{w_{ijk\bullet}} X_{ijkt} \quad \text{and} \quad w_{ijk\bullet} = \sum_{t=1}^{T_{ijk}} w_{ijkt} \qquad (58)$$

$$\begin{aligned}\hat{\Xi}_i^{(1)} &= z_i^{(1)}(X_{i\bullet\bullet\bullet}^{(zw)} - m)\\ &\quad - z_i^{(1)} \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{z_{ij}^{(12)}}{z_{i\bullet}^{(12)}} \frac{z_{ijk}^{(123)}}{z_{ij\bullet}^{(123)}} \left(\hat{\Xi}_j^{(2)} + \hat{\Xi}_k^{(3)} + \hat{\Xi}_{ik}^{(13)} + \hat{\Xi}_{jk}^{(23)}\right) \qquad (59)\end{aligned}$$

$$\begin{aligned}\hat{\Xi}_j^{(2)} &= z_j^{(2)}(X_{\bullet j\bullet\bullet}^{(zw)} - m)\\ &\quad - z_j^{(2)} \sum_{i=1}^{I}\sum_{k=1}^{K} \frac{z_{ij}^{(12)}}{z_{\bullet j}^{(12)}} \frac{z_{ijk}^{(123)}}{z_{ij\bullet}^{(123)}} \left(\hat{\Xi}_i^{(1)} + \hat{\Xi}_k^{(3)} + \hat{\Xi}_{ik}^{(13)} + \hat{\Xi}_{jk}^{(23)}\right) \qquad (60)\end{aligned}$$

$$\begin{aligned}\hat{\Xi}_k^{(3)} &= z_k^{(3)}(X_{\bullet\bullet k\bullet}^{(zw)} - m)\\ &\quad - z_k^{(3)} \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{z_{ik}^{(13)}}{z_{\bullet k}^{(13)}} \frac{z_{ijk}^{(123)}}{z_{i\bullet k}^{(123)}} \left(\hat{\Xi}_i^{(1)} + \hat{\Xi}_j^{(2)} + \hat{\Xi}_{ij}^{(12)} + \hat{\Xi}_{jk}^{(23)}\right) \qquad (61)\end{aligned}$$

$$\begin{aligned}\hat{\Xi}_{ij}^{(12)} &= z_{ij}^{(12)}(X_{ij\bullet\bullet}^{(zw)} - m)\\ &\quad - z_{ij}^{(12)} \sum_{k=1}^{K} \frac{z_{ijk}^{(123)}}{z_{ij\bullet}^{(123)}} \left(\hat{\Xi}_i^{(1)} + \hat{\Xi}_j^{(2)} + \hat{\Xi}_k^{(3)} + \hat{\Xi}_{ik}^{(13)} + \hat{\Xi}_{jk}^{(23)}\right) \qquad (62)\end{aligned}$$

$$
\begin{aligned}
\widehat{\Xi}_{ik}^{(13)} &= z_{ik}^{(13)}(X_{i\bullet k\bullet}^{(zw)} - m) \\
&\quad - z_{ik}^{(13)} \sum_{j=1}^{J} \frac{z_{ijk}^{(123)}}{z_{i\bullet k}^{(123)}} \left(\widehat{\Xi}_i^{(1)} + \widehat{\Xi}_j^{(2)} + \widehat{\Xi}_k^{(3)} + \widehat{\Xi}_{ij}^{(12)} + \widehat{\Xi}_{jk}^{(23)}\right)
\end{aligned} \tag{63}
$$

$$
\begin{aligned}
\widehat{\Xi}_{jk}^{(23)} &= z_{jk}^{(23)}(X_{\bullet jk\bullet}^{(zw)} - m) \\
&\quad - z_{jk}^{(23)} \sum_{i=1}^{I} \frac{z_{ijk}^{(123)}}{z_{\bullet jk}^{(123)}} \left(\widehat{\Xi}_i^{(1)} + \widehat{\Xi}_j^{(2)} + \widehat{\Xi}_k^{(3)} + \widehat{\Xi}_{ij}^{(12)} + \widehat{\Xi}_{ik}^{(13)}\right)
\end{aligned} \tag{64}
$$

The credibility factors appearing in (57) and in the terms in (57) are given by:

$$
z_i^{(1)} = \frac{z_{i\bullet}^{(12)}}{z_{i\bullet}^{(12)} + b^{(12)}/b^{(1)}}, \qquad z_{ij}^{(12)} = \frac{z_{ij\bullet}^{(123)}}{z_{ij\bullet}^{(123)} + b^{(123)}/b^{(12)}},
$$

$$
z_{ijk}^{(123)} = \frac{w_{ijk\bullet}}{w_{ijk\bullet} + s^2/b^{(123)}}.
$$

Finally, the credibility weighted means are defined in the usual manner, namely

$$
X_{ij\bullet\bullet}^{(zw)} = \sum_{k=1}^{K} \frac{z_{ijk}^{(123)}}{z_{ij\bullet}^{(123)}} X_{ijk}^{(w)} \quad \text{with} \quad z_{ij\bullet}^{(123)} = \sum_{k=1}^{K} z_{ijk}^{(123)}
$$

and

$$
X_{i\bullet\bullet\bullet}^{(zw)} = \sum_{j=1}^{J} \frac{z_{ij}^{(12)}}{z_{i\bullet}^{(12)}} X_{ij\bullet\bullet}^{(zw)} \quad \text{with} \quad z_{i\bullet}^{(12)} = \sum_{k=1}^{K} z_{ij}^{(12)}.
$$

The insured's credibility premium given in (57) is thus equal to the collective mean m plus two adjustment terms: the sum of the credibility estimators of the random variables $\Xi_i^{(1)}, \ldots, \Xi_{jk}^{(23)}$ and the collective mean. These $\Xi_i^{(1)}, \ldots, \Xi_{jk}^{(23)}$ are not given explicitly, but rather as the solution of a linear system of six equations. Note that the quantity $\Xi_{ijk}^{(123)}$ does not appear in the credibility premium formula.

8.3 Estimation of the Structure Parameters

Once again the credibility premium depends on the unknown structure parameters. The following estimators of the structure parameters are derived from those of Dannenburg (1995). As the crossed classification credibility model is fairly new, much work remains in the area of parameter estimation. For example, the variance estimators below can be negative and have no known optimality properties. Simulations have also demonstrated that they may have a large coefficient of variation.

A simple unbiased estimator of the collective mean m is the portfolio weighted average of the observations using the natural weights:

$$\hat{m} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{t=1}^{T_{ijk}} \frac{w_{ijkt}}{w_{\bullet\bullet\bullet\bullet}} X_{ijkt} \quad \text{where} \quad w_{\bullet\bullet\bullet\bullet} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{t=1}^{T_{ijk}} w_{ijkt} \tag{65}$$

An unbiased estimator of the parameter s^2 is simply

$$\hat{s}^2 = \frac{1}{d}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{t=1}^{T_{ijk}} w_{ijkt}(X_{ijkt} - X_{ijk\bullet}^{(w)})^2, \tag{66}$$

where d is equal to the total number of terms in the sum less the number of estimated means X_{ijkw}. For example, if the amount of experience is the same for each insured, that is if $T_{ijk} = T$ for all i, j, k, then $d = IJK(T-1)$.

Estimators of the variance components $b^{(1)}, \ldots, b^{(123)}$ can be given as solutions of a linear system of equations. First, we have,

$$\begin{aligned}
\sum_{i=1}^{I}\sum_{j=1}^{J} \tilde{g}_{ij}^{(12)} &\left(\sum_{k=1}^{K} \frac{w_{ijk\bullet}}{w_{ij\bullet}} (X_{ijk\bullet}^{(w)} - X_{ij\bullet\bullet}^{(w)})^2 - \frac{(K-1)}{w_{ij\bullet\bullet}} \hat{s}^2 \right) \\
= &\left(b^{(3)} + b^{(13)} + b^{(23)} + b^{(123)}\right) \\
&\times \left[1 - \sum_{i=1}^{I}\sum_{j=1}^{J} \tilde{g}_{ij}^{(12)} \sum_{k=1}^{K} \left(\frac{w_{ijk\bullet}}{w_{ij\bullet\bullet}} \right)^2 \right],
\end{aligned} \tag{67}$$

$$\sum_{i=1}^{I} \tilde{g}_i^{(1)} \left(\sum_{j=1}^{J} \sum_{k=1}^{K} \frac{w_{ijk\bullet}}{w_{i\bullet\bullet\bullet}} (X_{ijk\bullet}^{(w)} - X_{i\bullet\bullet\bullet}^{(w)})^2 - \frac{(JK-1)}{w_{i\bullet\bullet\bullet}} \hat{s}^2 \right)$$

$$\begin{aligned}
&= (b^{(2)} + b^{(12)}) \left[1 - \sum_{i=1}^{I} \tilde{g}_i^{(1)} \sum_{j=1}^{J} \left(\frac{w_{ij\bullet\bullet}}{w_{i\bullet\bullet\bullet}} \right)^2 \right] \\
&\quad + (b^{(3)} + b^{(13)}) \left[1 - \sum_{i=1}^{I} \tilde{g}_i^{(1)} \sum_{k=1}^{K} \left(\frac{w_{i\bullet k\bullet}}{w_{i\bullet\bullet\bullet}} \right)^2 \right] \\
&\quad + (b^{(23)} + b^{(123)}) \left[1 - \sum_{i=1}^{I} \tilde{g}_i^{(1)} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\frac{w_{ijk\bullet}}{w_{i\bullet\bullet\bullet}} \right)^2 \right] \qquad (68)
\end{aligned}$$

and

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{w_{ijk\bullet}}{w_{\bullet\bullet\bullet\bullet}} (X_{ijk\bullet}^{(w)} - X_{i\bullet\bullet\bullet}^{(w)})^2 - \frac{(IJK-1)}{w_{\bullet\bullet\bullet\bullet}} \hat{s}^2$$

$$\begin{aligned}
&= b^{(1)} \left[1 - \sum_{i=1}^{I} \left(\frac{w_{i\bullet\bullet\bullet}}{w_{\bullet\bullet\bullet\bullet}} \right)^2 \right] + b^{(2)} \left[1 - \sum_{j=1}^{J} \left(\frac{w_{\bullet j\bullet\bullet}}{w_{\bullet\bullet\bullet\bullet}} \right)^2 \right] \\
&\quad + b^{(3)} \left[1 - \sum_{k=1}^{K} \left(\frac{w_{\bullet\bullet k\bullet}}{w_{\bullet\bullet\bullet\bullet}} \right)^2 \right] + b^{(12)} \left[1 - \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\frac{w_{ij\bullet\bullet}}{w_{\bullet\bullet\bullet\bullet}} \right)^2 \right] \\
&\quad + b^{(13)} \left[1 - \sum_{i=1}^{I} \sum_{k=1}^{K} \left(\frac{w_{i\bullet k\bullet}}{w_{\bullet\bullet\bullet\bullet}} \right)^2 \right] + b^{(23)} \left[1 - \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\frac{w_{\bullet jk\bullet}}{w_{\bullet\bullet\bullet\bullet}} \right)^2 \right] \\
&\quad + b^{(123)} \left[1 - \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\frac{w_{ijk\bullet}}{w_{\bullet\bullet\bullet\bullet}} \right)^2 \right], \qquad (69)
\end{aligned}$$

where the $\tilde{g}^{(\cdot)}$s are arbitrary weights that sum to one. Note that the $\tilde{g}^{(\cdot)}$s are not necessarily non-negative weights. Equal weights should be used if, a priori, no risk factor appears more important than the others for parameter estimation. Otherwise, the natural weights can be used.

With the appropriate permutations of the order of summation, one can then generate as many equations as there are unknowns (in this case there will be seven equations). Specifically, the first equation above should be summed in the order ijk, then ikj, and then jki. For the second equation, the summation orders are ijk, jik, and kij.

The estimation procedure is more straightforward than it initially appears. However, it requires much computational resources as the number of risk factors and categories increases.

For the sake of illustration, the entire system of equations that needs to be solved to find estimators of $b^{(1)}, \dots, b^{(123)}$ is given below with equal weights $\tilde{g}$. First the following terms are defined:

$$x_1 = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\sum_{k=1}^{K} \frac{w_{ijk\bullet}}{w_{ij\bullet\bullet}} (X_{ijk\bullet}^{(w)} - X_{ij\bullet\bullet}^{(w)})^2 - \frac{(K-1)}{w_{ij\bullet\bullet}} \hat{s}^2 \right),$$

$$x_2 = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} \left(\sum_{j=1}^{J} \frac{w_{ijk\bullet}}{w_{i\bullet k\bullet}} (X_{ijk\bullet}^{(w)} - X_{i\bullet k\bullet}^{(w)})^2 - \frac{(J-1)}{w_{i\bullet k\bullet}} \hat{s}^2 \right),$$

$$x_3 = \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\sum_{i=1}^{I} \frac{w_{ijk\bullet}}{w_{\bullet jk\bullet}} (X_{ijk\bullet}^{(w)} - X_{\bullet jk\bullet}^{(w)})^2 - \frac{(I-1)}{w_{\bullet jk\bullet}} \hat{s}^2 \right),$$

$$x_4 = \frac{1}{I} \sum_{i=1}^{I} \left(\sum_{j=1}^{J} \sum_{k=1}^{K} \frac{w_{ijk\bullet}}{w_{i\bullet\bullet\bullet}} (X_{ijk\bullet}^{(w)} - X_{i\bullet\bullet\bullet}^{(w)})^2 - \frac{(JK-1)}{w_{i\bullet\bullet\bullet}} \hat{s}^2 \right),$$

$$x_5 = \frac{1}{J} \sum_{j=1}^{J} \left(\sum_{i=1}^{I} \sum_{k=1}^{K} \frac{w_{ijk\bullet}}{w_{\bullet j\bullet\bullet}} (X_{ijk\bullet}^{(w)} - X_{\bullet j\bullet\bullet}^{(w)})^2 - \frac{(IK-1)}{w_{\bullet j\bullet\bullet}} \hat{s}^2 \right),$$

$$x_6 = \frac{1}{K} \sum_{k=1}^{K} \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{w_{ijk\bullet}}{w_{\bullet\bullet k\bullet}} (X_{ijk\bullet}^{(w)} - X_{\bullet\bullet k\bullet}^{(w)})^2 - \frac{(IJ-1)}{w_{\bullet\bullet k\bullet}} \hat{s}^2 \right),$$

$$x_7 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{w_{ijk\bullet}}{w_{\bullet\bullet\bullet\bullet}} (X_{ijk\bullet}^{(w)} - X_{\bullet\bullet\bullet\bullet}^{(w)})^2 - \frac{(IJK-1)}{w_{\bullet\bullet\bullet\bullet}} \hat{s}^2$$

and

$$a_1 = 1 - \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\frac{w_{ijk\bullet}}{w_{ij\bullet\bullet}} \right)^2, \qquad a_2 = 1 - \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} \sum_{j=1}^{J} \left(\frac{w_{ijk\bullet}}{w_{ik\bullet}} \right)^2,$$

$$a_3 = 1 - \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{I} \left(\frac{w_{ijk\bullet}}{w_{\bullet jk\bullet}} \right)^2, \qquad a_4 = 1 - \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\frac{w_{ij\bullet\bullet}}{w_{i\bullet\bullet\bullet}} \right)^2,$$

$$a_5 = 1 - \frac{1}{I} \sum_{i=1}^{I} \sum_{k=1}^{J} \left(\frac{w_{ik\bullet}}{w_{i\bullet\bullet\bullet}} \right)^2, \qquad a_6 = 1 - \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\frac{w_{ijk\bullet}}{w_{i\bullet\bullet\bullet}} \right)^2,$$

$$a_7 = 1 - \frac{1}{J} \sum_{j=1}^{J} \sum_{i=1}^{I} \left(\frac{w_{ij\bullet\bullet}}{w_{\bullet j\bullet\bullet}} \right)^2, \qquad a_8 = 1 - \frac{1}{J} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\frac{w_{\bullet jk\bullet}}{w_{\bullet j\bullet\bullet}} \right)^2,$$

$$a_9 = 1 - \frac{1}{J}\sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{k=1}^{K}\left(\frac{w_{ijk\bullet}}{w_{\bullet j\bullet\bullet}}\right)^2, \qquad a_{10} = 1 - \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{I}\left(\frac{w_{ik\bullet}}{w_{\bullet\bullet k\bullet}}\right)^2,$$

$$a_{11} = 1 - \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{J}\left(\frac{w_{\bullet jk\bullet}}{w_{\bullet\bullet k\bullet}}\right)^2, \qquad a_{12} = 1 - \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{w_{ijk\bullet}}{w_{\bullet\bullet k\bullet}}\right)^2,$$

$$a_{13} = 1 - \sum_{i=1}^{I}\left(\frac{w_{i\bullet\bullet\bullet}}{w_{\bullet\bullet\bullet\bullet}}\right)^2, \qquad a_{14} = 1 - \sum_{j=1}^{J}\left(\frac{w_{\bullet j\bullet\bullet}}{w_{\bullet\bullet\bullet\bullet}}\right)^2,$$

$$a_{15} = 1 - \sum_{k=1}^{K}\left(\frac{w_{\bullet\bullet k\bullet}}{w_{\bullet\bullet\bullet\bullet}}\right)^2, \qquad a_{16} = 1 - \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{w_{ij\bullet\bullet}}{w_{\bullet\bullet\bullet\bullet}}\right)^2,$$

$$a_{17} = 1 - \sum_{i=1}^{I}\sum_{k=1}^{K}\left(\frac{w_{ik\bullet}}{w_{\bullet\bullet\bullet\bullet}}\right)^2, \qquad a_{18} = 1 - \sum_{j=1}^{J}\sum_{k=1}^{K}\left(\frac{w_{\bullet jk\bullet}}{w_{\bullet\bullet\bullet\bullet}}\right)^2,$$

$$a_{19} = 1 - \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\frac{w_{ijk\bullet}}{w_{\bullet\bullet\bullet\bullet}}\right)^2.$$

The estimators $\hat{b}^{(1)}, \ldots, \hat{b}^{(123)}$ are then the solutions of the system of linear equations

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = \begin{pmatrix} 0 & 0 & a_1 & 0 & a_1 & a_1 & a_1 \\ 0 & a_2 & 0 & a_2 & 0 & a_2 & a_2 \\ a_3 & 0 & 0 & a_3 & a_3 & 0 & a_3 \\ 0 & a_4 & a_5 & a_4 & a_5 & a_6 & a_6 \\ a_7 & 0 & a_8 & a_7 & a_9 & a_8 & a_9 \\ a_{10} & a_{11} & 0 & a_{12} & a_{10} & a_{11} & a_{12} \\ a_{13} & a_{14} & a_{15} & a_{16} & a_{17} & a_{18} & a_{19} \end{pmatrix} \begin{pmatrix} \hat{b}^{(1)} \\ \hat{b}^{(2)} \\ \hat{b}^{(3)} \\ \hat{b}^{(12)} \\ \hat{b}^{(13)} \\ \hat{b}^{(23)} \\ \hat{b}^{(123)} \end{pmatrix}$$

9 Closing Comments

One of the objectives of this paper is to distinguish limited fluctuation credibility from greatest accuracy credibility. The historical remarks emphasize that the former approach was developed to determine a level above which the experience of an insured would be considered fully credible. The latter evolved from a desire to estimate an insured's risk premium as precisely as possible; its success depended on the wider acceptance of Bayesian statistics among statisticians and actuaries. The hierarchical model is a generalization of the Bühlmann-Straub model and other one-level (or one-risk factor) credibility models.

This model can prove especially useful in practice when the actuary is faced with a very large portfolio. Finally, the recent crossed classification credibility model was introduced. This model relies upon variance components models and mostly generalizes the hierarchical model.

Credibility theory is sometimes called "the mathematics of heterogeneity." As such, the models presented here—or the others cited in references—could be used in numerous fields outside the traditional ones. So far, credibility theory has mostly been used in some areas of casualty insurance or group life insurance. However, one could think of many situations, in actuarial science or not, where the key problem is heterogeneity of the data. Credibility theory in these other fields could constitute a useful tool, just as it has been to actuaries for almost a century now.

References

Bailey, A.L. "A Generalized Theory of Credibility." *Proceedings of the Casualty Actuarial Society* 32 (1945): 13-20.

Bailey, A.L. "Credibility Procedures, Laplace's Generalization of Bayes' Rule and the Combination of Collateral Knowledge with Observed Data." *Proceedings of the Casualty Actuarial Society* 37 (1950): 7-23.

Bühlmann, H. "Experience Rating and Credibility." *ASTIN Bulletin* 4 (1967): 199-207.

Bühlmann, H. "Experience Rating and Credibility." *ASTIN Bulletin* 5 (1969): 157-165.

Bühlmann, H. and Straub, E. "Glaubgwürdigkeit für Schadensätze." *Bulletin of the Swiss Association of Actuaries* 70 (1970): 111-133. English translation by C.E. Brooks.

Bühlmann, H. and Jewell, W.S. "Hierarchical Credibility Revisited." *Bulletin of the Swiss Association of Actuaries* 87 (1987): 35-54.

Dannenburg, D. "Crossed Classification Credibility Models." *Transactions of the 25th International Congress of Actuaries* 4 (1995): 1-35.

Dannenburg, D.R., Kaas, R., and Goovaerts, M.J. *Practical Actuarial Credibility Models.* Amsterdam, Holland: Institute of Actuarial Science and Econometrics, University of Amsterdam, 1996.

De Vylder, F. "Geometrical Credibility." *Scandinavian Actuarial Journal* (1976): 121-149.

De Vylder, F. "Optimal Semilinear Credibility." *Bulletin of the Swiss Association of Actuaries* 76 (1976): 27-40.

De Vylder, F. "Parameter Estimation in Credibility Theory." *ASTIN Bulletin* 10 (1978): 99-112.

De Vylder, F. "Practical Credibility Theory with Emphasis on Parameter Estimation." *ASTIN Bulletin* 12 (1981): 115-131.

De Vylder, F. "Practical Models in Credibility Theory, Including Parameter Estimation." In *Premium Calculation in Insurance.* NATO ASI Series, (1984): 133-150.

De Vylder, F. *Advanced Risk Theory.* Brussels, Belgium: Éditions de l'Université de Bruxelles, Swiss Association of Actuaries, 1996.

De Vylder, F. and Goovaerts, M.J. "Estimation of the Heterogeneity Parameter in the Bühlmann-Straub Credibility Theory Model." *Insurance: Mathematics and Economics* 10 (1991): 233-238.

De Vylder, F. and Goovaerts, M.J. "Optimal Parameter Estimation Under Zero-Excess Assumptions in a Bühlmann-Straub Model." *Insurance: Mathematics and Economics* 11 (1992): 167-171.

Dubey, A. and Gisler, A. "On Parameter Estimation in Credibility." *Bulletin of the Swiss Association of Actuaries* 81 (1981): 187-211.

Feller, W. "On a General Class of 'Contagious' Distributions." *Annals of Mathematical Statistics* 14 (1943): 389-400.

Feller, W. *An Introduction to Probability Theory and its Application,* Volume 2. New York, N.Y.: Wiley, 1966.

Fischer, A. "Discussion of Whitney (1918)." *Proceedings of the Casualty Actuarial Society* 5 (1919): 139-145.

Gerber, H.U. *An Introduction to Mathematical Risk Theory.* Philadelphia, Pa.: Huebner Foundation, 1979.

Gisler, A. "Optimal Trimming of Data in the Credibility Model." *Bulletin of the Swiss Association of Actuaries* 80 (1980): 313-325.

Gisler, A. and Reinhard, P. "Robust Credibility." *ASTIN Bulletin* 23 (1993): 117-143.

Goel, P.K. "On the Implications of Credible Means Being Exact Bayesian." *Scandinavian Actuarial Journal* (1982): 41-46.

Goovaerts, M.J. and Hoogstad, W.J. *Credibility Theory. Surveys of Actuarial Studies,* No. 4. Rotterdam, Holland: Nationale-Nederlanden N.V., 1987.

Goovaerts, M.J., Kaas, R., van Heerwaarden, A.E., and Bauwelinckx, T. *Effective Actuarial Methods.* Amsterdam, Holland: North-Holland, 1990.

Goulet, V. "On Approximations in Limited Fluctuation Credibility." *Proceedings of the Casualty Actuarial Society* 84 (1997): 533-552.

Hachemeister, C.A. "Credibility for Regression Models with Application to Trend." In *Credibility, Theory and Applications, Proceedings of the Berkeley Actuarial Research Conference on Credibility.* New York, N.Y.: Academic Press, 1975: 129-163.

Hogg, R.V. and Craig, A.T. *Introduction to Mathematical Statistics,* (4th edition). New York, N.Y.: Macmillan, 1978.

Hogg, R.V. and Klugman, S.A. *Loss Distributions.* New York, N.Y.: Wiley, 1984.

Jewell, W.S. "Credible Means Are Exact Bayesian for Exponential Families." *ASTIN Bulletin* 8 (1974): 77-90.

Jewell, W.S. "The Use of Collateral Data in Credibility Theory: A Hierarchical Model." *Giornale dell'Istituto Italiano degli Attuari* 38 (1975): 1-16.

Keffer, R. "An Experience Rating Formula." *Transactions of the Actuarial Society of America* 30 (1929): 130-139.

Künsch, H.R. "Robust Methods for Credibility." *ASTIN Bulletin* 22 (1992): 33-49.

Longley-Cook, L.H. "An Introduction to Credibility Theory." *Proceedings of the Casualty Actuarial Society* 49 (1962): 194-221.

Mayerson, A.L. "A Bayesian View of Credibility." *Proceedings of the Casualty Actuarial Society* 51 (1964): 85-104.

Mowbray, A.H. "How Extensive a Payroll Exposure is Necessary to Give a Dependable Pure Premium?" *Proceedings of the Casualty Actuarial Society* 1 (1914): 25-30.

Norberg, R. "The Credibility Approach to Ratemaking." *Scandinavian Actuarial Journal* (1979): 181-221.

Norberg, R. "Empirical Bayes Credibility." *Scandinavian Actuarial Journal* (1980): 177-194.

Norberg, R. "Hierarchical Credibility: Analysis of a Random Effects Linear Model with Nested Classification." *Scandinavian Actuarial Journal* (1986): 204-222.

Philbrick, S.W. "An Examination of Credibility Concepts." *Proceedings of the Casualty Actuarial Society* 68 (1981): 195-219.

Robbins, H. "An Empirical Bayes Approach to Statistics." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability,* Volume 30. Berkeley, Calif.: University of California Press, 1955: 157-163.

Robbins, H. "An Empirical Bayes Approach to Statistics." *Annals of Mathematical Statistics* 30 (1964): 1-20.

Russell, B. *Human Knowledge: Its Scope and Limits.* London, England: George Allen and Unwin Ltd, 1948.

Searle, S.R. *Linear Models.* New York, N.Y.: Wiley, 1971.

Searle, S.R., Casella, G. and McCulloch, C.E. *Variance Components.* New York, N.Y.: Wiley, 1992.

Sundt, B. "A Hierarchical Credibility Regression Model." *Scandinavian Actuarial Journal* (1979): 107-114.

Sundt, B. "A Multi-Level Hierarchical Credibility Regression Model." *Scandinavian Actuarial Journal* (1980): 25-32.

Sundt, B. "On Greatest Accuracy Credibility with Limited Fluctuation." *Scandinavian Actuarial Journal* (1992): 109-119.

Whitney, A.W. "The Theory of Experience Rating." *Proceedings of the Casualty Actuarial Society* 4 (1918): 275-293.

Actuarial Techniques in Risk Pricing and Cash Flow Analysis for U.K. Bank Loans

Philip Booth* and Duncan E.P. Walsh†

Abstract‡

A cash flow model is developed to set the price for a loan to a borrower with known risks. Similarities are noted between this model and those used for profit testing in life insurance. We emphasize aspects that reasonably can be treated in several ways and also indicate where the cash flow model differs from the pricing methods usually employed in bank lending. The sensitivity of interest rates to various parameters of the model such as the length of loan and the expected default rate is examined. Also, we examine how features of loans, including cash back and early repayments, can be priced.

Key words and phrases: *credit risk, default rate, equity, expenses, mortgages, net present value*

*Philip Booth is a senior lecturer in actuarial science at City University. He has previously worked in the investment department at Axa Equity and Law and has taught in Eastern Europe. He is the co-author of two books and a number of papers in actuarial science. Mr. Booth is an assistant editor of the *British Actuarial Journal* and is a member of a number of Institute of Actuaries' Committees.

Mr. Booth's address is: Department of Actuarial Science and Statistics, City University, Northampton Square, London ECIV OHB, ENGLAND. Internet address: *p.booth@city.ac.uk*

†Duncan E.P. Walsh, Ph.D., is a researcher in actuarial science at City University. He joined the department in 1996 after completing a Diploma in Actuarial Science at City University. Dr. Walsh obtained his bachelor's degree in mathematics and astronomy from the University of Leicester, England, and his Ph.D. in astrophysical sciences from Princeton University, USA. He was an astrophysicist, focusing on cosmology, serving as a researcher at Cambridge University (England) and at Hokkaido University (Japan). His current research interests include long-term care and property investments.

Dr. Walsh's address is: Department of Actuarial Science and Statistics, City University, Northampton Square, London ECIV OHB, ENGLAND. Internet address: *depw@city.ac.uk*

‡This paper is based on a previous paper entitled "An Actuarial Approach to the Pricing of Banking Risk" in *Proceedings of the 1997 Faculty and Institute of Actuaries Investment Conference*, Faculty and Institute of Actuaries, 1998 (forthcoming). The authors wish to thank the following: Iain Allan for helpful comments and practical guidance, the Institute of Actuaries and the Royal Bank of Scotland for funding of research grants, the anonymous referees, and the editor for their help in "Americanizing" the paper.

1 Introduction

The principal objective of a bank is to make loans in such a manner as to provide its shareholders with a healthy return on their equity capital. To this end, banks make large corporate loans, small corporate loans, personal loans (including mortgages and auto loans and unsecured loans) and operate credit cards. For the large loans there is more information required (e.g., financial statements and accounts and the institution's credit rating). Risk pricing, whereby different interest rates are set according to the default risk associated with the loan, is accepted as the normal market practice for setting loan rates. For smaller corporate loans and personal loans there is some credit risk information, both on the economic background and individual risk. The normal market practice for these loans, however, is to charge a uniform price to those who are offered loans. The risk analysis merely determines the decision of whether to lend or not; it does not affect the interest rate charged.

There are several types of risks banks face including:

- Credit risk, i.e., the risk that some borrowers will default; it is a bank's major consideration in the lending process;[1]
- Market risk, i.e., the risk of changes in the market value of assets;
- Liquidity risk, i.e., the risk of not holding enough liquid assets as the bank's liabilities are predominantly short term in nature; and
- Operational risk, i.e., fraud, computer failure, terrorism, etc.

We are concerned primarily with credit risk and its impact on smaller corporate loans and personal loans. The risk of default for these loans is a major consideration because there is often not much relevant information on the borrower's ability to repay the loan. The market, liquidity, and operational risks are discussed by Allan et al., (1998).

Once the potential borrower's credit risk is known,[2] the bank can choose to decline or accept the request for a loan. If the request is acceptable, the bank must decide at what level to set two key parameters:

[1] As banks are aware of the possibility of loan defaults, they make an annual provision for the resulting bad debts, typically 1 percent of outstanding loans. This figure varies as bad debt is sensitive to the state of the economy. Values for the provisions for loans to various industries are given by Davis (1993). More recent ratios are given in the *Banking Act Report* (annual) and the *Annual Abstract of Banking Statistics* (but these do not include industry breakdowns).

[2] A discussion of how credit risk is determined is given in Appendix A.

the interest rate charged on the loan and the amount of capital set aside to back the loan.

The bank can calculate the minimum interest rate required to provide sufficient returns to the bank and then compare this with the current market rate of interest for this type of loan. Though an increased interest rate will raise the expected proceeds from the loan, it may also increase the borrower's default risk.

The capital allocation is based on two considerations: regulatory and economic. First, there is a regulatory requirement for banks to hold a certain amount of capital to protect the bank from insolvency.[3] In the U.K. banks generally have held capital of around 10 percent of these assets, 6 percent of which has been equity capital.[4] Within each category (e.g., commercial loans or mortgages) the regulatory capital requirement includes no allowance for differences in default risk.[5]

Second, there is a general preference among shareholders for a stable pattern of returns. Variable credit losses can lead to variable returns. This variability of returns can be reduced by holding more capital, as any losses will lead to a smaller percentage loss of capital. The economic capital requirement increases with the variability of credit risk.

Increasing the amount of capital that supports a loan reduces the expected return on capital, however, unless there is an increase in interest rates. Thus the interest rate to charge on the loan and the amount of capital set aside to back the loan are interdependent.

2 An Overview of Basic Cash Flow Models

Cash flow models for bank loans have a variety of uses such as: (i) calculating the return on equity capital to see whether lending is likely to be profitable at a particular interest rate, (ii) examining the impact of various parameters on default scenarios, and (iii) examining the cost to

[3]Banks are required under the Basle Accord to hold capital of at least 8 percent of their risk-weighted assets, including at least 4 percent equity capital; the remainder will be debt capital. Risk-weighted assets include 100 percent of commercial loans, 50 percent of mortgages, and 0 percent of government debt.

[4]The amount of capital held varies with time, e.g., both quantities increased through the first half of the 1990s, and varies between banks, with some banks holding total capital of 14 percent and equity capital of 9 percent.

[5]This means that loans to large corporations need as much capital backing (per £ of loan) as do loans to individuals. This contrasts with risk-adjusted or economic capital that takes risk into account. It is a concern among banks that this encourages high risk lending, as it is inefficient to hold large amounts of capital for low risk loans.

the bank of loan features such as guaranteeing fixed interest rates or allowing early repayments. In considering the makeup of a cash flow model, however, we will focus on calculations of expected net present value and return on equity capital.

Cash flow equations in bank lending may be complicated, but the ideas are not different from those used in other actuarial cash flow models. There are terms for the amount of inflow and outflow, the timing of these flows, the probability that they occur, and a discount factor. The complexity arises because there are many parties to consider (the shareholders, the borrower, the bank's treasury, and the providers of debt capital). In addition there is a possibility of premature termination of the loan by default or early repayment, both of which yield income (including default recoveries and surrender fees).

A cash flow model can be based on the total amount of loans outstanding and other directly linked quantities such as capital, monthly expenses, monthly net interest income, losses due to defaults, and fees from early repayments. These quantities may be fixed or variable.

There are three approaches to examining cash flows relating to lending: the cohort loan approach, steady state portfolio approach, and the whole business approach. These approaches have different uses and they do not give the same value for the profitability of a particular class of business.

Cohort Loan Approach: Only the income and outgo relating to one or a group (cohort) of similar loans issued at the same time are considered.

Steady State Portfolio Approach: Here lending is viewed as a steady state process whereby at any time a given block of loans is outstanding, it is supported by a proportional amount of capital. The outstanding loans give rise to streams of interest payments and expenses. The development of individual loans is ignored for such calculations.[6] For a bank that already has many loans written and expects to both issue new loans and receive final payments on others at a steady rate, it is not necessary to consider each loan in detail. (Although it is probably useful to consider a set of loans from start-up when pricing.)

Whole Business Approach: We consider the whole business of lending including (i) the costs of establishing computer systems, training

[6]A variant is when the loan book is expected to fluctuate, but the entire set of loans still is considered rather than each loan. This is a simpler, more practical approach to the analysis of loan cash flows than studying each loan. It omits some details relating to the timing of payments.

staff, and so on; (ii) a model of the growth rate of the business; and (iii) all of the cash flows arising directly from the lending. This differs from the other approaches by including more expenses (not just those directly related to marketing and maintaining loans).

The cohort loan valuation method can be inappropriate when the arrangements for repaying the funds for the loans and for paying the expenses generated by the loan are based on the portfolio of loans. In this case, it would be possible to use the proportional repayment calculations in the individual loan cash flows to handle the funding costs, but it would still be necessary to make decisions regarding what portion of the net income generated by a particular loan in each month is to be paid to the providers of capital and what portion is to be used to meet expenses of the portfolio.

The steady state method cannot readily be used for pricing new business or considering the profitability of a new type of loan. It is best to use individual loans to assess the value of features such as initial discounts or early repayments, because the timing of payments is crucial in this instance. When looking at the whole portfolio, income generated now is compared with the cost of capital in place now rather than being matched with the capital that was invested in the past to back the loans that are now generating income.

When deciding on the profitability of a new line of business (for example, personal loans sold by telephone), the whole business approach may be better as calculating the value of each loan is not sufficient. There will be substantial start-up costs and marketing expenses may be higher per loan arranged in the first year compared with loans made later. These extra costs must be spread across all loans of this class made over a period of several years. A cash flow analysis must include these initial costs, estimates of the growth in volume of lending (e.g., quarterly estimates for the first five years), and the income and outgo pertaining to each loan.

As we are primarily considering interest rate setting and the profitability of a tranche of loans, we will use the cohort loan approach.[7]

---

[7] A note on the words used in this paper: *cohort* and *tranche* are used when describing loans issued at the same time; *steady state*, *portfolio*, and *book* are used to describe a combination of loans at different stages of development. The phrase *set of loans* is used for either of these two situations, i.e., it is an alternative to using the plural *loans*.

3 Cash Flow Model for a Cohort of Loans

The cash flow model is developed sequentially. First we consider only the loan and the equity capital, with expenses, debt capital, defaults, and early repayments being ignored. These items are introduced singly later in the paper.

3.1 Two Sources of Funds

Each cash flow resulting from a loan can be split into two sources: flows that belong to the shareholders and flows that do not belong to the shareholders. To assess the profitability of a loan it is essential to correctly identify from which source each element of a cash flow came. This idea is developed further in the following example, with expenses ignored for simplicity. Let

r_F = Cost of funds, which is at least the money market rate and possibly larger to allow for the expenses of the treasury department;

r_L = Interest received on the loan;

r_C = Interest earned by the bank's equity capital; and

C_t = Net cash flow at t.

To make a one year loan of say, 100, at $r_L = 12$ percent, the bank's lending department will, in turn, have to borrow the same amount of money from the bank's treasury department. The bank's treasury department in turn will acquire the money from retail deposits or short-term borrowing in the wholesale markets. The treasury will charge the lending department a rate $r_F = 10$ percent for the use of this money. It is the two percent difference between r_F and r_L percent that is the crucial element in the profitability calculations.[8]

The bank also must set aside equity capital of 5 percent of the loan to back each loan. These funds will be invested in the money markets and earn $r_C = 8$ percent during the year. Depending on how the bank's treasury operates, r_C could be equal to r_F.

Thus, as far as the shareholders are concerned, the initial cash flow is $C_0 = -5$, i.e., the capital set aside. The end of year cash flow is

[8]In the cohort approach, the global weighted average margin on all loans would be determined so that it was sufficient to provide an appropriate return on capital. There would be insufficient explicit consideration given to the cash flow pertaining to individual loans.

$$C_1 = 100(1 + r_L) - 100(1 + r_F) + 5(1 + r_C) = 112 - 110 + 5.4 = 7.4.$$

Note that these cash flows that belong to the shareholders are small in comparison with the total cash flows that occur in the lending process.

The profitability of this loan is related to the net present value (*NPV*) which is given by

$$NPV(r) = C_0 + \frac{C_1}{1 + r}$$

where r an interest rate. A loan is profitable if $NPV(r_H) > 0$ where r_H is the hurdle rate.[9] The internal rate of return (*IRR*) is the rate of interest that solves $NPV(IRR) = 0$. In most cases where *IRR* is greater than the hurdle rate the project will be sufficiently profitable. In this example, with a (pre-tax) hurdle rate of 20 percent, we have $NPV(0.20) = 1.17$ and an *IRR* of 48 percent. Thus, with no expenses or defaults, it is sufficiently profitable to lend with these rates of interest.

In practice a more complicated method may be used to determine if a loan is sufficiently profitable. This method involves (i) calculating *NPV* at above the hurdle rate, with a check that this is positive; (ii) calculating *IRR*, with a check to see that it is sufficiently high; and (iii) a check on the sensitivity of *NPV* to relevant variables.

The implications of having two sources have been detailed because, although splitting cash flow may be obvious, this situation is not mentioned in standard business finance texts in discussing the valuation of cash flows. One method mentioned in texts is to compare the *IRR* of all the flows with an average of the returns required by those involved with the project (here the providers of capital and the bank's treasury). This example gives a combined initial outgo of 105 and the final income of 117.4, yielding

$$NPV(r) = -105 + \frac{117.4}{1 + r}$$

and $IRR = 11.81$ percent. Some authors have suggested that *NPV* could be calculated using a discount rate equal to a weighted average of the

[9]The hurdle rate is set by the bank according to the riskiness of the loan using a risk versus return model such as the capital asset pricing model. It is higher than the rate of interest charged by the treasury because the treasury is exposed to less risk than the loan department. The treasury has a prior claim on any income; if there is any shortfall (e.g., because of a loan default) capital, if available, will be used to make up the difference.

two rates of interest involved. This is called the *weighted average cost of capital approach.*[10] This method is not as precise as considering the flows to and from each participant separately.

The separation of a cash flow into several streams is familiar to actuaries—for example in the context of unit-linked life policies (e.g., Squires, 1986) where premium income is split between a unit fund (belonging to the policyholder) and a sterling fund (belonging to the office). A closer analogy to the two sources of funds required in bank lending is where a negative sterling fund is used in a life office (e.g., Hare and McCutcheon, 1991). In such a situation the initial strain caused by establishing a policy is partly backed by capital, which requires one rate of interest, and partly by internal funds, which require a lower rate of interest.

3.2 The Basic Mathematical Model

We begin with a basic cash flow model that consists of loan repayments from the borrower to the loan department and from the loan department to the treasury.

In general, most personal loans or mortgages are amortized over time by level installments that include both interest and principal elements. An alternative approach is to use a sinking fund arrangement where a series of interest only payments are made and a final complete repayment of the principal. The sinking fund approach has capital outstanding for a longer period and therefore may have a greater risk to the bank than the amortization approach. In the amortization situation there is also a release of capital each month, as the capital requirement is likely to be proportional to the amount of the loan outstanding.

The amortization method is used throughout this paper. Without loss of generality, we assume loans are repaid on a monthly basis. The following notations are needed:

[10] See, for example, Higson 1986, Chapter 16, or Brealey and Myers, 1991, Chapter 19.

$$
\begin{aligned}
L_0 &= \text{Size of loan;} \\
X &= \text{Size of the level monthly installment to amortized } L_0; \\
L_t &= \text{Loan outstanding at end of month } t,\ t = 1, 2, \ldots; \\
K_0 &= \text{Initial capital;} \\
n &= \text{Duration of loan in months;} \\
i_L &= \text{Monthly interest rate on loan;} \\
i_F &= \text{Monthly interest rate on funds;} \\
i_C &= \text{Monthly interest rate on set aside capital;} \\
i_H &= \text{Monthly hurdle rate; and} \\
v_H &= 1/(1 + i_H).
\end{aligned}
$$

Note that throughout this paper the symbol r is the annual percentage rate (APR) corresponding to i. So, for example, $(1 + i_L)^{12} = 1 + r_L$.

It is well known that X and L_t are given by:

$$
X = \frac{L_0}{a_{\overline{n}|\, i_L}} \tag{1}
$$

$$
L_t = X a_{\overline{n-t}|\, i_L} \tag{2}
$$

where $a_{\overline{n}|\, i}$ is the present value of an annuity of one per month paid in arrears for n months evaluated at interest rate i.[11]

Let B_t denote the amount paid to the bank's treasury at the end of month t. Two possible schemes are considered for B_t, a uniform scheme and a proportional scheme. These schemes lead to

$$
B_t = \begin{cases} L_0 / a_{\overline{n}|\, i_F} & \text{Uniform Scheme;} \\ (1 + i_F)L_{t-1} - L_t & \text{Proportional Scheme.} \end{cases} \tag{3}
$$

The uniform repayment scheme involves n equal payments to the treasury; this is the same pattern as the initial intended payments by the borrower to the bank. The proportional scheme assumes that, at the start of each month, the bank borrows an amount equal to the loan outstanding at the time and repays this with interest at the cost of funds rate at the end of the month. It implicitly assumes that it will be

[11]As this paper does not focus on risks relating to changes in base rates, these formulae for X and L_t have been based on a constant interest rate throughout the term of the loan.

possible, throughout the term of a long loan for the treasury to be able to borrow the amount of money that already has been lent by the bank. Note that if a loan is repaid early, the uniform repayment plan ignores this while the proportional plan adapts by bringing forward the return of money to the treasury.

Both patterns have advantages: the uniform method fixes in advance the interest paid on the borrowed funds (the margin over base rate is fixed) so that uncertainty about future movements in these interest rates can be removed from the lending decision. The treasury knows in advance the pattern of payments it will receive from the lending department. The proportional method ensures that the amount borrowed at any time is the same as the amount being lent.

For the remainder of this paper we will use the proportional repayment method as this equates more closely to procedures followed in practice.[12]

Some more notation is required for cash flow modeling:

$$\begin{aligned}
K_t &= \text{Equity capital outstanding at end of month } t; \\
&= \frac{K_0 a_{\overline{n-t|}\, i_L}}{a_{\overline{n|}\, i_L}}; && (4) \\
RTK_t &= \text{Equity capital returned at the of month } t; \\
&= K_{t-1} - K_t; && (5) \\
i_C K_{t-1} &= \text{Interest earned on equity capital during month } t.
\end{aligned}$$

The release of capital implied by these definitions matches the repayments of principal by the borrower; thus the amount of capital is kept in proportion to the loan outstanding. This procedure should not be followed if analysis suggests that the loan is becoming more risky. The capital backing the lending should be kept at a level sufficient to cover future losses.

Using the basic model, the net monthly income (*NMI*) and net present value of the loan, from the viewpoint of the shareholders, is given by

$$NMI_t = X - [(1 + i_F)L_{t-1} - L_t] + i_C K_{t-1} + RTK_t \qquad (6)$$

$$NPV(i_H) = -K_0 + \sum_{t=1}^{n} NMI_t v_H^t. \qquad (7)$$

[12]Appendix B contains a discussion of the differences arising under the uniform repayment pattern and includes an explanation of why what appears to be a bookkeeping choice is more important to the derived profitability of a loan than are real features such as default rates.

Note that this value depends on the decision on how to arrange the repayments to the bank's treasury, the withdrawal of capital each month, and the borrower's repayment pattern.

Example 1

Tables 1 and 2 illustrate the development of the various terms in the cash flow equation using the following parameters: $L_0 = 5,000$, $K_0 = 0.05L_0$, $n = 36$ months, $r_H = 20$ percent, $r_L = 12$ percent, $r_F = 10$ percent, and $r_C = 8$ percent.

Table 1
Cash Flows in Respect of Borrower's Repayments

Month	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	5000.000	47.444	117.166	164.610	39.871	117.166	157.037
2	4882.834	46.332	118.278	164.610	38.936	118.278	157.215
3	4764.555	45.210	119.401	164.610	37.993	119.401	157.394
4	4645.155	44.077	120.533	164.610	37.041	120.533	157.575
5	4524.621	42.933	121.677	164.610	36.080	121.677	157.757
6	4402.944	41.779	122.832	164.610	35.110	122.832	157.941
7	4280.112	40.613	123.997	164.610	34.130	123.997	158.128
8	4156.115	39.437	125.174	164.610	33.141	125.174	158.315
9	4030.941	38.249	126.362	164.610	32.143	126.362	158.505
10	3904.580	37.050	127.561	164.610	31.136	127.561	158.696
11	3777.019	35.839	128.771	164.610	30.118	128.771	158.890
12	3648.248	34.617	129.993	164.610	29.092	129.993	159.085
13	3518.255	33.384	131.226	164.610	28.055	131.226	159.281
14	3387.029	32.139	132.472	164.610	27.009	132.472	159.480
15	3254.557	30.882	133.729	164.610	25.952	133.729	159.681
20	2573.104	24.416	140.195	164.610	20.518	140.195	160.713
25	1858.701	17.637	146.974	164.610	14.822	146.974	161.795
30	1109.754	10.530	154.080	164.610	8.849	154.080	162.929
35	324.593	3.080	161.530	164.610	2.588	161.530	164.119
36	163.063	1.547	163.063	164.610	1.300	163.063	164.363

Notes: Column (1) = Loan at start of month; Column (2) = Interest paid by borrower; Column (3) = Return of principal by borrower; Column (4) = Total paid by borrower; Column (5) = Interest paid to treasury; Column (6) = Return of principal to treasury; and Column (7) = Total paid to treasury.

Table 2

Cash Flows to Capital

Month	(1)	(2)	(3)	(4)	(5)	(6)
1	250.000	1.609	5.858	15.040	0.985	14.813
2	244.142	1.571	5.914	14.881	0.970	14.435
3	238.228	1.533	5.970	14.719	0.955	14.064
4	232.258	1.494	6.027	14.557	0.941	13.698
5	226.231	1.456	6.084	14.393	0.927	13.340
6	220.147	1.416	6.142	14.227	0.913	12.987
7	214.006	1.377	6.200	14.060	0.899	12.641
8	207.806	1.337	6.259	13.891	0.886	12.301
9	201.547	1.297	6.318	13.720	0.872	11.967
10	195.229	1.256	6.378	13.548	0.859	11.639
11	188.851	1.215	6.439	13.374	0.846	11.316
12	182.412	1.174	6.500	13.199	0.833	10.999
13	175.913	1.132	6.561	13.022	0.821	10.688
14	169.351	1.090	6.624	12.843	0.808	10.382
15	162.728	1.047	6.686	12.663	0.796	10.082
20	128.655	0.828	7.010	11.735	0.738	8.660
25	92.935	0.598	7.349	10.762	0.684	7.361
30	55.488	0.357	7.704	9.742	0.634	6.176
35	16.230	0.104	8.077	8.673	0.588	5.096
36	8.153	0.052	8.153	8.453	0.579	4.892
Sum of Present Value of Monthly Cash Flows						336.50
Less Initial Capital Allocation						-250.00
Net Present Value of Loan						86.50

Notes: Column (1) = Capital at start of month; Column (2) = Interest earned on capital; Column (3) = Return of capital; Column (4) = Net cash flow at end of month; Column (5) = Discount factor; and Column (6) = Net present value.

The net present value is 86.5, the *IRR* is 54.16 percent, and the loan interest rate that would provide a zero *NPV* at a 20 percent hurdle rate is 10.58 percent. If the hurdle rate is 20 percent, 10.6 percent can be regarded as the minimum loan interest rate, ignoring expenses and defaults.

Table 1 shows the constant repayments (164.610) by the borrower split between decreasing interest payments and increasing principal repayments. The total amount paid to the treasury increases each month under the proportional repayment scheme. (Under the uniform repayment scheme the monthly payment to the treasury would be 160.326.) Table 2 shows that each month's net cash flow is positive, and the size decreases as the size of the loan reduces. The net cash flow is calculated using:

$$\begin{aligned}\text{Net cash flow} \quad = \quad & \text{Total paid by borrower} \\ & - \text{Total paid to treasury} \\ & + \text{Interest earned on capital} \\ & + \text{Return of capital.}\end{aligned}$$

3.3 Inclusion of Expenses

There are several ways to deal with expenses, particularly initial expenses, and these methods lead to different values for the profitability of a loan and different sensitivities of the return on equity capital to parameters such as default rate.

Expenses are incurred in establishing the loan, maintaining it, and closing it. The cash flow treatment for these three types of expense is best considered separately.

The initial expenses included in the loan pricing calculations refer only to the costs directly attributable to selling and establishing new loans. They do not include overhead costs or the costs of establishing a line of business. Initial expenses may be met (i) by borrowing from the treasury, (ii) by using equity capital, or (iii) from the net income of existing loans. These three methods are discussed below.

If the initial expenses are borrowed from the treasury they must be repaid, with interest, at some later time using the repayments received on the loan. One way of accounting for this is to amortize these expenses over the term of the loan (using the interest rate applying to the cost of funds); therefore, a portion of each loan repayment will be applied to these start-up costs. This is equivalent to the uniform method of repaying borrowed funds. An alternative is to use the equivalent of the proportional method for the repayment of funds. For example, if the initial expenses are 1 percent of the loaned amount, any future payments to the treasury will be increased 1 percent to allow for the cost of these expenses, including interest. This proportional method is adopted here.

The equations for NMI_t and NPV, including the proportional method of repayment of initial expenses, E_0, are:

$$
\begin{aligned}
NMI_t &= X - [(1+i_F)L_{t-1} - L_t]\frac{L_0 + E_0}{L_0} + i_C K_{t-1} + RTK_t && (8)\\
NPV(i_H) &= -K_0 + \sum_{t=1}^{n} NMI_t v_H^t. && (9)
\end{aligned}
$$

Initial expenses could be met from capital (excluding regulatory capital) on the grounds that there is a risk that they will not be recovered because the borrower fails to make sufficient payments. As default probabilities tend to decrease over the life of the loan, the likelihood of recovering these initial expenses is maximized if the first few installments paid by the borrower are used to meet the expenses rather than contribute to profit. (In life insurance profit testing the initial expenses generally are charged to capital.) But because initial expenses can be relatively large, it would require a substantial increase in the capital outlay for a loan if the expenses had to be met in this way. Moreover, this capital would be consumed immediately and therefore would not earn any interest. Hence, this would be a costly approach. It is also not an approach used in practice.

If the initial expenses are met by capital rather than borrowing, the E_0 term is not needed in equation (8) for net monthly income. The NMI and NPV equations must be changed to

$$
\begin{aligned}
NMI_t &= X - [(1+i_F)L_{t-1} - L_t] + i_C K_{t-1} + RTK_t && (10)\\
NPV(i_H) &= -K_0 - E_0 + \sum_{t=1}^{n} NMI_t v_H^t. && (11)
\end{aligned}
$$

As E_0 will not, in general, vary directly with L_0, NPV will not vary directly in proportion to L_0; thus, small loans will be unprofitable except at high interest rates.

3.4 Debt Capital

As well as holding equity capital, the bank will hold debt capital (also called tier 2 capital) as part of the regulatory requirements. Providers will require a return in excess of what the bank can earn by putting the money in the cash market. This generates an extra expense each month of ${}^{D}K_{t-1} \times (i_D - i_C)$ where ${}^{D}K_t$ is the amount of debt capital held at the end of month t, which will depend on the size of the loan outstanding, and i_D is the monthly interest which has to be paid on this debt.

These payments are included in the cash flow model in the same way as the running expenses, E_t:

$$\begin{aligned} NMI_t &= X - [(1+i_F)L_{t-1} - L_t]\frac{L_0 + E_0}{L_0} + i_C K_{t-1} + RTK_t \\ &\quad - E_t - {}^{D}K_{t-1}(i_D - i_C) \\ NPV(i_H) &= -K_0 + \sum_{t=1}^{n} NMI_t v_H^t. \end{aligned}$$

The interest payments relating to debt capital are important but changes in the amount of debt capital held do not alter the cash flows to or from the providers of equity capital.

Example 2

Table 3 and 4 show the behavior of the terms in this equation. The parameters are the same as for Example 1, but with the inclusion of the additional terms: $E_0 = 50$, $E_t = 0$, $r_D = 10$ percent, and ${}^{D}K_0 = K_0$ (i.e., the initial debt capital and equity capital are equal). Here the initial costs are met by borrowing from the treasury, and this produces a net present value of 36.20 (at a discount rate of 20 percent), an internal rate of return of 34.11 percent, and a break-even loan interest rate (at the 20 percent hurdle rate) of 11.41 percent. If capital were used for these initial costs the *NPV* would be 30.28, the *IRR* would be 29.60 percent, and the break-even loan rate would be 11.50 percent.

3.5 Loans Defaults

Some borrowers will default on the repayment of their loans. On some of the defaulted loans, the bank will be unable to recover the full amount of the outstanding principal resulting in a loss.

Table 3
Cash Flows in Respect of Borrower Allowing For Expenses

Month	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	5000.000	47.444	117.166	164.610	5050.000	40.269	118.338	158.607
2	4882.834	46.332	118.278	164.610	4931.662	39.326	119.461	158.787
3	4764.555	45.210	119.401	164.610	4812.201	38.373	120.595	158.968
4	4645.155	44.077	120.533	164.610	4691.606	37.412	121.739	159.150
5	4524.621	42.933	121.677	164.610	4569.868	36.441	122.894	159.335
6	4402.944	41.779	122.832	164.610	4446.974	35.461	124.060	159.521
7	4280.112	40.613	123.997	164.610	4322.914	34.472	125.237	159.709
8	4156.115	39.437	125.174	164.610	4197.676	33.473	126.426	159.898
9	4030.941	38.249	126.362	164.610	4071.251	32.465	127.625	160.090
10	3904.580	37.050	127.561	164.610	3943.625	31.447	128.836	160.283
11	3777.019	35.839	128.771	164.610	3814.789	30.420	130.059	160.478
12	3648.248	34.617	129.993	164.610	3684.730	29.383	131.293	160.675
13	3518.255	33.384	131.226	164.610	3553.438	28.336	132.539	160.874
14	3387.029	32.139	132.472	164.610	3420.899	27.279	133.796	161.075
15	3254.557	30.882	133.729	164.610	3287.103	26.212	135.066	161.278
20	2573.104	24.416	140.195	164.610	2598.835	20.723	141.597	162.320
25	1858.701	17.637	146.974	164.610	1877.288	14.970	148.443	163.413
30	1109.754	10.530	154.080	164.610	1120.852	8.938	155.621	164.559
35	324.593	3.080	161.530	164.610	327.839	2.614	163.146	165.760
36	163.063	1.547	163.063	164.610	164.694	1.313	164.694	166.007

Notes: Column (1) = Loan at start of month; Column (2) = Interest paid by borrower; Column (3) = Return of principal by borrower; Column (4) = Total paid by borrower; Column (5) = Amount owed to treasury at start; Column (6) = Interest paid to treasury; Column (7) = Return of principal to treasury; and Column (8) = Total paid to treasury.

Table 4
Cash Flows to Capital Allowing For Expenses

Month	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	250.000	1.609	5.858	0.385	13.085	0.985	12.887
2	244.142	1.571	5.914	0.376	12.932	0.970	12.545
3	238.228	1.533	5.970	0.367	12.779	0.955	12.209
4	232.258	1.494	6.027	0.358	12.623	0.941	11.879
5	226.231	1.456	6.084	0.348	12.467	0.927	11.555
6	220.147	1.416	6.142	0.339	12.308	0.913	11.236
7	214.006	1.377	6.200	0.330	12.149	0.899	10.923
8	207.806	1.337	6.259	0.320	11.988	0.886	10.616
9	201.547	1.297	6.318	0.310	11.825	0.872	10.314
10	195.229	1.256	6.378	0.301	11.661	0.859	10.017
11	188.851	1.215	6.439	0.291	11.495	0.846	9.726
12	182.412	1.174	6.500	0.281	11.327	0.833	9.439
13	175.913	1.132	6.561	0.271	11.158	0.821	9.158
14	169.351	1.090	6.624	0.261	10.988	0.808	8.882
15	162.728	1.047	6.686	0.251	10.816	0.796	8.611
20	128.655	0.828	7.010	0.198	9.930	0.738	7.328
25	92.935	0.598	7.349	0.143	9.001	0.684	6.156
30	55.488	0.357	7.704	0.085	8.027	0.634	5.089
35	16.230	0.104	8.077	0.025	7.006	0.588	4.117
36	8.153	0.052	8.153	0.013	6.796	0.579	3.933
Sum of Present Value of Monthly Cash Flows							286.20
Less Initial Capital Allocation							-250.00
Net Present Value of Loan							36.20

Notes: Column (1) = Capital at start of month; Column (2) = Interest earned on capital; Column (3) = Return of capital; Column (4) = Net interest on debt capital; Column (5) = Net cash flow at end of month; Column (6) = Discount factor; and Column (7) = Net present value.

The entire loss consists, however, of the unrecovered outstanding principal plus any previous missed interest payments plus any extra expenses incurred in the collection of the loan. Hence, it is possible for the entire loss to exceed the outstanding principal. Thus both the frequency of default and the resulting losses are crucial factors in the pricing of loans.

The notation for the cash flow model requires the following additions:

$$\begin{aligned}
q_t &= \text{Probability of loan default during month } t;\\
P_t^{(q)} &= \text{Probability that loan remains in effect at end of month } t;\\
&= \prod_{j=1}^{t}(1-q_j);\\
f_t &= \text{Expected ratio of the loss during month } t \text{ to } L_t.
\end{aligned}$$

The q_t and f_t must be estimated in advance, perhaps from historical data relating to similar loans. The estimation of these rates, however, is a major challenge.

The expected loss during month t is $q_t \times P_{t-1}^{(q)} \times f_t \times L_{t-1}$. This formulation of default recovery assumes either that the recoveries are made immediately or that $(1-f_t)L_{t-1}$ refers to the present value at time t of the amounts recovered at later dates.

The expected net monthly income of the loan, taking account of the defaults, becomes:

$$\begin{aligned}
NMI_t &= XP_t^{(q)} - [(1+i_F)P_{t-1}^{(q)}L_{t-1} - P_t^{(q)}L_t]\frac{L_0+E_0}{L_0}\\
&\quad + i_C\, P_{t-1}^{(q)}K_{t-1} + (P_{t-1}^{(q)}K_{t-1} - P_t^{(q)}K_t)\\
&\quad - E_t - P_{t-1}^{(q)\,D}K_{t-1}(i_D - i_C) + q_t P_{t-1}^{(q)}(1-f_t)L_{t-1}. \qquad (12)
\end{aligned}$$

Even with all of the features that have been incorporated in the cash flow model, the complexity of the situation is understated because loans will not be split between on-going and defaulted. There are likely to be some loans in arrears, for which provisions may be set aside before the default date (or the date of successful repayment of the amount owed). For mortgages a loan can be in arrears for more than two years before the situation is resolved, so this is not just a small matter of detail.

Example 3

This example continues from Example 2 with the inclusion of two new parameters: q_t = 0.2 percent (per month); and f_t = 0.2 (i.e., 80 percent of the outstanding loan is recovered). The columns in Tables 5 and 6 with asterisks have been explicitly adjusted by survival probabilities. (Some of the other columns are sums and thereby acquire the adjustment indirectly.)

The loan is only just profitable at a discount rate of 20 percent with an *NPV* of 1.24. The internal rate of return is 20.47 percent, and the break-even loan interest rate at the 20 percent hurdle rate is 11.98 percent. The total money received from continuing borrowers is less than the amount paid each month to the treasury, and the recovery of a substantial portion of each defaulted loan is an important component of the net monthly income.

Table 7 displays the net present values and internal rates of return using the same parameters as used to construct Tables 5 and 6, but with monthly default rates included. Table 7 demonstrates that the *IRR* is somewhat more variable when the initial expenses are met by borrowing rather than being met from capital.

3.6 Early Repayment of Loans

The terms of a loan sometimes will, for a fee, allow the borrower to repay the loan early. Early repayments can be an important feature of long-term loans such as mortgages where many borrowers move or may switch banks in search of the lowest interest rates. The bank cannot rely on the receipt of a full number of interest payments to provide the required profits. The problem is amplified by the fact that U.K. mortgage loans often include a reduced interest rate in the first year and by the concentration of expenses and default risks near the beginning of the loan period. The bank relies on later interest payments to make lending worthwhile.[13]

The loan is only just profitable at a discount rate of 20 percent with an *NPV* of 1.24. The internal rate of return is 20.47 percent, and the break-even loan interest rate at the 20 percent hurdle rate is 11.98 percent.

[13]In the U.S.A., the problem of early repayment is dealt with mainly through the use of points, i.e., interest paid in advance at start of loan for a reduced interest rate. This reduces the early repayment risk.

Table 5
Cash Flows in Respect of Borrower Allowing For Defaults

Month	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	0.998	5000.000	47.349	116.932	164.281	5050.000	40.269	128.201	168.471
2	0.996	4882.834	46.147	117.806	163.953	4931.662	39.247	128.827	168.074
3	0.994	4764.555	44.939	118.686	163.625	4812.201	38.220	129.458	167.678
4	0.992	4645.155	43.725	119.572	163.297	4691.606	37.188	130.095	167.282
5	0.990	4524.621	42.506	120.465	162.971	4569.868	36.150	130.737	166.887
6	0.988	4402.944	41.280	121.365	162.645	4446.974	35.108	131.384	166.492
7	0.986	4280.112	40.048	122.272	162.320	4322.914	34.060	132.037	166.097
8	0.984	4156.115	38.810	123.185	161.995	4197.676	33.007	132.695	165.702
9	0.982	4030.941	37.566	124.105	161.671	4071.251	31.949	133.359	165.308
10	0.980	3904.580	36.315	125.032	161.348	3943.625	30.885	134.029	164.914
11	0.978	3777.019	35.059	125.966	161.025	3814.789	29.817	134.704	164.521
12	0.976	3648.248	33.796	126.907	160.703	3684.730	28.743	135.385	164.128
13	0.974	3518.255	32.526	127.855	160.381	3553.438	27.663	136.072	163.735
14	0.972	3387.029	31.251	128.810	160.061	3420.899	26.578	136.764	163.342
15	0.970	3254.557	29.968	129.772	159.741	3287.103	25.487	137.463	162.950
20	0.961	2573.104	23.457	134.692	158.150	2598.835	19.950	141.043	160.993
25	0.951	1858.701	16.776	139.799	156.574	1877.288	14.267	144.775	159.042
30	0.942	1109.754	9.916	145.099	155.015	1120.852	8.434	148.665	157.098
35	0.932	324.593	2.872	150.599	153.471	327.839	2.442	152.718	155.160
36	0.930	163.063	1.440	151.724	153.164	164.694	1.224	153.549	154.773

Notes: Column (1) = Probability of payment at end of the month; Column (2) = Loan at start of month; Column (3) = Interest paid by borrower; Column (4) = Return of principal by borrower; Column (5) = Total paid by borrower; Column (6) = Amount owed to treasury at start; Column (7) = Interest paid to treasury; Column (8) = Return of principal to treasury; and Column (9) = Total paid to treasury.

Table 6

Cash Flows to Capital Allowing For Defaults

Month	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	250.000	1.609	6.347	0.385	8.000	11.380	0.985	11.209
2	244.142	1.568	6.378	0.375	7.797	11.245	0.970	10.909
3	238.228	1.527	6.409	0.365	7.593	11.109	0.955	10.614
4	232.258	1.485	6.440	0.356	7.388	10.973	0.941	10.326
5	226.231	1.444	6.472	0.346	7.182	10.836	0.927	10.043
6	220.147	1.402	6.504	0.336	6.975	10.699	0.913	9.766
7	214.006	1.360	6.536	0.326	6.766	10.560	0.899	9.495
8	207.806	1.318	6.569	0.316	6.557	10.422	0.886	9.229
9	201.547	1.276	6.602	0.305	6.347	10.282	0.872	8.968
10	195.229	1.234	6.635	0.295	6.136	10.142	0.859	8.713
11	188.851	1.191	6.669	0.285	5.923	10.002	0.846	8.462
12	182.412	1.148	6.702	0.275	5.710	9.861	0.833	8.217
13	175.913	1.105	6.736	0.264	5.496	9.719	0.821	7.977
14	169.351	1.062	6.771	0.254	5.280	9.577	0.808	7.742
15	162.728	1.018	6.805	0.244	5.063	9.433	0.796	7.511
20	128.655	0.797	6.982	0.191	3.963	8.709	0.738	6.427
25	92.935	0.570	7.167	0.136	2.834	7.967	0.684	5.449
30	55.488	0.337	7.360	0.081	1.675	7.208	0.634	4.569
35	16.230	0.098	7.560	0.023	0.485	6.430	0.588	3.778
36	8.153	0.049	7.601	0.012	0.243	6.273	0.579	3.630
Sum of Present Value of Monthly Cash Flows								251.24
Less Initial Capital Allocation								-250.00
Net Present Value of Loan								1.24

Notes: Column (1) = Capital at start of month; Column (2) = Interest earned on capital; Column (3) = Return of capital; Column (4) = Net interest on debt capital; Column (5) = Recovery from defaulted loans; Column (6) = Net cash flow at end of month; Column (7) = Discount factor; and Column (8) = Net present value.

Table 7
Impact of Expenses on *NPV* and *IRR*
With $E_0 = 50$, $L_0 = 5,000$, and $K_0 = 250$

Monthly Default Rate	Expenses Borrowed From Treasury		Expenses Paid From Capital	
	NPV	*IRR*	*NPV*	*IRR*
0.0%	36.20	34.11%	30.28	29.60%
0.2%	1.24	20.47%	-4.56	18.58%
0.4%	-32.26	8.06%	-37.94	8.40%
0.6%	-64.37	-3.21%	-69.92	-1.00%

The early termination of a loan can be put in a cash flow model in a similar manner to defaults. Let G_t denote the fee charged for early repayment in month t; and R_t denote the probability that a loan that has survived to the end of month t is repaid at that time. The survival probability for a loan becomes

$$P_t^{(qr)} = \prod_{j=1}^{t}(1 - q_j)(1 - R_j). \tag{13}$$

The proportion of the original loans that default at the end of month t will be $q_t P_{t-1}^{(qr)}$, while the repayments will be $R_t(1 - q_t)P_{t-1}^{(qr)}$. The expression for net monthly income becomes:

$$\begin{aligned} NMI_t \quad = \quad & P_{t-1}^{(qr)} X(1 - q_t) - [(1 + r_F)P_{t-1}^{(qr)} L_{t-1} - P_t^{(qr)} L_t]\frac{L_0 + E_0}{L_0} - E_t \\ & + i_C P_{t-1}^{(qr)} K_{t-1} + (P_{t-1}^{(qr)} K_{t-1} - P_t^{(qr)} K_t) \\ & - (i_D - i_C)^D K_{t-1} P_{t-1}^{(qr)} + q_t P_{t-1}^{(qr)} (1 - f_t) L_{t-1} \\ & + R_t(1 - q_t)P_{t-1}^{(qr)} (L_t + G_t). \end{aligned} \tag{14}$$

The approach we have taken here is an interesting contrast to the approach taken in Allan et al., (1998). That paper assumes that all loans survive the average period of seven years for a U.K. mortgage and then are repaid. Pricing is set so that the average loan provides an appropriate profit. In the U.K. this method would provide reasonable results and would provide similar results to this system where a distribution of future repayment times is used. If there were less inertia in the loan market, a more active fee structure may need to be developed to penalize early repayers or a probability distribution of repayment times may need to be used to estimate the expected cost and variability of cost of early repayment.

Example 4

This example builds on Tables 1 through 6 with the inclusion of early repayments. The two new parameter values are $G_t = 0.01L_t$ and

$$R_t = \begin{cases} 0 & t \leq 12 \\ 0.002 & t > 12. \end{cases}$$

There are four extra columns in Tables 8 and 9: the probability of a loan surviving to the start of the month (i.e., $P_{t-1}^{(qr)}$), the probability of early repayment, the amount of early repayments, and the fees accompanying these repayments.

Because there are no early repayments in the first year of this example, the first twelve months are identical to Example 3. Thereafter, *NMI* is initially greater than in the no repayment example, but in the last months of the loan it is less than in Example 3. *NPV* of 1.52 is marginally higher than without repayments, indicating that the 1 percent fee is sufficient to cover the loss of later positive cash flows. Other values for this loan include an internal rate of return of 20.58 percent and a break-even interest rate of 11.97 percent.

Table 8

Cash Flows in Respect of Borrower Allowing For Prepayments

Month	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	1.000	0.998	5000.0	47.3	116.9	164.3	5050.0	40.3	128.2	168.5
2	0.998	0.996	4882.8	46.1	117.8	164.0	4931.7	39.2	128.8	168.1
3	0.996	0.994	4764.6	44.9	118.7	163.6	4812.2	38.2	129.5	167.7
4	0.994	0.992	4645.2	43.7	119.6	163.3	4691.6	37.2	130.1	167.3
5	0.992	0.990	4524.6	42.5	120.5	163.0	4569.9	36.2	130.7	166.9
6	0.990	0.988	4402.9	41.3	121.4	162.6	4447.0	35.1	131.4	166.5
7	0.988	0.986	4280.1	40.0	122.3	162.3	4322.9	34.1	132.0	166.1
8	0.986	0.984	4156.1	38.8	123.2	162.0	4197.7	33.0	132.7	165.7
9	0.984	0.982	4030.9	37.6	124.1	161.7	4071.3	31.9	133.4	165.3
10	0.982	0.980	3904.6	36.3	125.0	161.3	3943.6	30.9	134.0	164.9
15	0.968	0.967	3254.6	29.8	129.3	159.1	3287.1	25.4	143.0	168.4
20	0.949	0.947	2573.1	23.1	132.8	155.9	2598.8	19.7	143.7	163.4
25	0.930	0.929	1858.7	16.4	136.5	152.9	1877.3	13.9	144.5	158.5
30	0.912	0.910	1109.8	9.6	140.2	149.8	1120.9	8.2	145.4	153.6
35	0.894	0.892	324.6	2.7	144.1	146.9	327.8	2.3	146.4	148.8
36	0.890	0.889	163.1	1.4	144.9	146.3	164.7	1.2	146.6	147.8

Notes: Column (1) = Probability of loan surviving to start of month; Column (2) = Probability of payment at end of the month; Column (3) = Loan at start of month; Column (4) = Interest paid by borrower; Column (5) = Return of principal by borrower; Column (6) = Total paid by borrower; Column (7) = Amount owed to treasury at start; Column (8) = Interest paid to treasury; Column (9) = Return of principal to treasury; and Column (10) = Total paid to treasury.

Table 9
Cash Flows to Capital Allowing For Prepayments

Month	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1	250.0	1.6	6.3	0.4	8.0	0.000	0.000	0.000	11.380	0.985	11.209
2	244.1	1.6	6.4	0.4	7.8	0.000	0.000	0.000	11.245	0.970	10.909
3	238.2	1.5	6.4	0.4	7.6	0.000	0.000	0.000	11.109	0.955	10.614
4	232.3	1.5	6.4	0.4	7.4	0.000	0.000	0.000	10.973	0.941	10.326
5	226.2	1.4	6.5	0.3	7.2	0.000	0.000	0.000	10.836	0.927	10.043
6	220.1	1.4	6.5	0.3	7.0	0.000	0.000	0.000	10.699	0.913	9.766
7	214.0	1.4	6.5	0.3	6.8	0.000	0.000	0.000	10.560	0.899	9.495
8	207.8	1.3	6.6	0.3	6.6	0.000	0.000	0.000	10.422	0.886	9.229
9	201.5	1.3	6.6	0.3	6.3	0.000	0.000	0.000	10.282	0.872	8.968
10	195.2	1.2	6.6	0.3	6.1	0.000	0.000	0.000	10.142	0.859	8.713
15	162.7	1.0	7.1	0.2	5.0	0.002	6.033	0.060	9.697	0.796	7.721
20	128.7	0.8	7.1	0.2	3.9	0.002	4.610	0.046	8.818	0.738	6.507
25	92.9	0.6	7.2	0.1	2.8	0.002	3.179	0.032	7.937	0.684	5.428
30	55.5	0.3	7.2	0.1	1.6	0.002	1.740	0.017	7.054	0.634	4.472
35	16.2	0.1	7.2	0.0	0.5	0.002	0.291	0.003	6.168	0.588	3.624
36	8.2	0.0	7.3	0.0	0.2	0.002	0.000	0.000	5.990	0.579	3.467
Sum of Present Value of Monthly Cash Flows											251.24
Less Initial Capital Allocation											-250.00
Net Present Value of Loan											1.24

Notes: Column (1) = Capital at start of month; Column (2) = Interest earned on capital; Column (3) = Return of capital; Column (4) = Net interest on debt capital; Column (5) = Recovery from defaulted loans; Column (6) = Probability of early repayment; Column (7) = Early repayments; Column (8) = Early repayment fees; Column (9) = Net cash flow at end of month; Column (10) = Discount factor; and Column (11) = Net present value.

3.7 Parameter Dependence

Equation (14) is used to generate Table 10, which shows how the net present value (at a 20 percent hurdle rate), the internal rate of return, and the break-even loan rate (also at a 20 percent hurdle rate) change as the various inputs of the cash flow model are altered. These three values are given for the case where initial expenses are met by borrowing and the case where capital is used for these expenses. The standard model has the following parameters: $L_0 = 5,000$, $K_0/L_0 = 0.05$, $n = 36$ months, $r_L = 12$ percent, $r_F = 10$ percent, $r_C = 8$ percent, $r_H = 20$ percent, $E_0 = 50$, ${}^DK_0/L_0 = 0.05$, $r_D = 10$ percent, $E_t = 0$, $q_t = 0.2$ percent, $f_t = 0.2$, $G_t/L_t = 1$ percent, and $R_t = 0$ for $t \leq 12$ and $= 0.2$ percent for $t > 12$. All other entries in Table 10 differ only in one value. This standard model is the one used in Example 4.

The following observations may be drawn from Table 10:

- In all cases, *NPV* is greater when initial expenses are paid by borrowing rather than by using capital. (The same discount rate has been used for the two scenarios though it may be reasonable to use a lower rate when capital is used.) Clearly the option of borrowing from the treasury is cheaper than using capital to finance expenses. The borrowing option, however, would lead to greater variability of returns on a smaller amount of capital.

- In terms of the increase in break-even interest rate, the effect of the choice between these two methods of paying for the initial expenses is sensitive to the size of the loan, the hurdle rate used, and the amount of initial expenses.

- The loan rate and the cost of funds are more important than the interest rate earned on set aside capital and the interest paid on debt capital. The lending margin between the loan rate and cost of funds dwarfs all other cash flows to shareholders. Therefore the profitability is likely to be heavily dependent on the interest margin.

- The profitability is less sensitive to the hurdle rate than it is to the loan interest rate or the cost of funds. (The loan rate and the cost of fund rates are varied independently in the above table, hence the margin on the loan is changed.)

Table 10

The Impact of Changes in Various Inputs of the Cash Flow Model

Parameter Changes			Borrowed			Capital		
	Old	New	*NPV*	*IRR*(%)	r_L(%)	*NPV*	*IRR*(%)	r_L(%)
Standard			1.52	20.58	11.97	-4.24	18.67	12.07
L_0	5,000	1,000	-35.08	-54.74	14.97	-40.85	-17.34	15.46
	5,000	3000	-16.78	9.33	12.47	-22.54	9.51	12.64
	5,000	10000	47.28	29.05	11.60	41.52	27.16	11.65
K_0/L_0	0.05	0.01	29.42	105.25	11.50	23.66	46.54	11.60
	0.05	0.25	18.96	36.08	11.68	13.20	27.54	11.78
	0.05	0.10	-33.35	13.96	12.57	-39.11	13.58	12.66
n	36	18	-21.66	5.89	12.66	-24.86	6.60	12.76
	36	60	26.97	27.04	11.69	18.40	23.93	11.79
r_L	0.12	0.11	-57.59	-0.37	11.97	-63.34	1.39	12.07
	0.12	0.13	60.59	44.98	11.97	54.80	38.28	12.07
r_F	0.10	0.08	123.75	75.92	9.92	116.78	62.01	10.04
	0.10	0.11	-58.83	-0.70	13.00	-63.99	1.27	13.09
r_C	0.08	0.07	-4.60	18.26	12.08	-10.37	16.78	12.18
	0.08	0.09	7.60	22.93	11.87	1.84	20.58	11.97
r_H	0.20	0.15	15.53	20.58	11.75	12.46	18.67	11.80
	0.20	0.25	-10.94	20.58	12.19	-19.09	18.67	12.34
	0.20	0.30	-22.08	20.58	12.40	-32.37	18.67	12.59

Notes: The original parameter values are given at the start of Section 3.7 and are displayed in the column labeled "OLD" for convenience. The parameter changes are from the column labeled "OLD" to the one labeled "NEW".

Table 10 (Cont.)
The Impact of Changes in Various Inputs of the Cash Flow Model

Parameter Changes			Borrowed			Capital		
	Old	New	*NPV*	*IRR*(%)	r_L(%)	*NPV*	*IRR*(%)	r_L(%)
Standard			1.52	20.58	11.97	-4.24	18.67	12.07
E_0	50	0	45.76	37.58	11.22	45.76	37.58	11.22
	50	100	-42.71	3.70	12.72	-54.24	5.63	12.92
${}^{D}K_0/L_0$	0.05	0.025	4.55	21.74	11.92	-1.21	19.62	12.02
	0.05	0.10	-4.53	18.29	12.08	-10.29	16.80	12.17
r_D	0.10	0.09	4.54	21.74	11.92	-1.23	19.61	12.02
	0.10	0.11	-1.46	19.44	12.02	-7.23	17.74	12.12
E_t	0	1	-25.99	10.66	12.44	-31.76	10.05	12.54
q_t	0.002	0.000	36.30	34.22	11.40	30.41	29.69	11.50
	0.002	0.004	-31.80	8.17	12.55	-37.44	8.49	12.65
f_t	0.2	0.1	17.24	26.72	11.71	11.48	23.66	11.81
	0.2	0.5	-45.62	3.66	12.77	-51.39	4.75	12.87
	0.2	1.0	-124.20	20.09	14.12	-129.96	-15.35	14.22
G_t/L_t	0.01	0.00	0.94	20.36	11.98	-4.83	18.49	12.08
	0.01	0.02	2.11	20.80	11.96	-3.65	18.86	12.06
$R_t, t > 12$	0.2	0.0	1.24	20.47	11.98	-4.56	18.58	12.08
	0.2	0.4	1.80	20.69	11.97	-3.93	18.76	12.07

Notes: The original parameter values are given at the start of Section 3.7 and are displayed in the column labeled "OLD" for convenience. The parameter changes are from the column labeled "OLD" to the one labeled "NEW".

- Because expenses are a higher proportion of small loans, larger loans are more profitable than small loans, both in absolute terms and per unit of capital deployed, all other things being equal. This suggests that differential interest rates with loan size and/or loan fees would be an appropriate charging policy.
- Similarly, long loans produce more profit than short loans because there is a longer time over which to amortize initial expenses.
- The amount of equity capital is more significant than the amount of debt capital because equity capital requires a higher return.
- Both initial expenses and running expenses are important.
- An extra expense of 1 per month on a loan of 5,000 requires the interest rate to be raised 0.5 percent.
- Doubling initial expenses (to 100 per loan of 5,000) would cause a greater loss than doubling the default rates (suggesting that there is a limit to the expense that should be used to assess the default risk of the borrowers).
- Though the default rate is relevant to profitability, the effect of doubling the default rate is no worse than halving the duration of the loan, starting from the parameters of the standard loan.
- Halving the loan loss fraction has a similar impact to halving the loan default rate. This is not surprising, as both parameters relate to the expected loss from a loan.
- With the parameters explored here, early repayments and associated fees are not important.

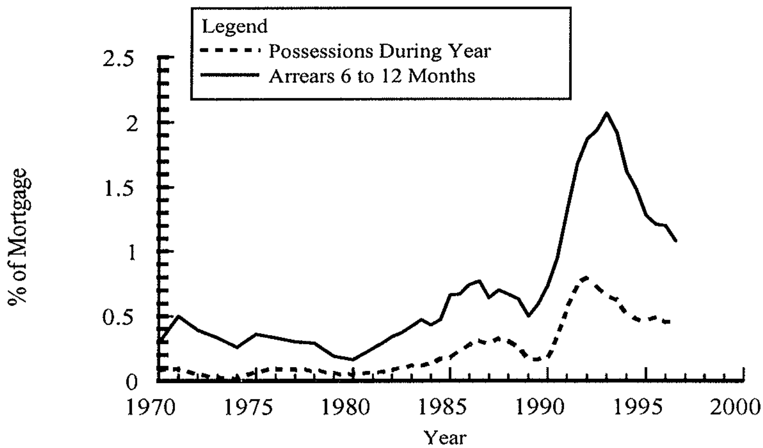
4 Variability of Default Rates and Costs

Data on mortgage arrears and possessions have been collected since 1969 by the Building Societies Association (published in the *BSA Bulletin*) and later the Council of Mortgage Lenders (published in *Housing Finance*).[14] The data suggest evidence of cyclical behavior in the 1970s in the proportions of mortgages ending in possession; see Figure 1. The

[14] A mortgage is said to be in *arrears* whenever at least one scheduled monthly payment is not paid by a certain date. A *mortgage possession* (also called a *repossession*) occurs when the mortgage is in arrears and the bank thus terminates the mortgage and takes ownership of the house. This usually requires a court order.

proportions rose throughout the first half of the 1980s, peaking in the first half of 1987 with an annualized rate of 0.33 percent of mortgaged properties taken into possession. Between the first half of 1989 and the second half of 1991 the annualized rate rose from 0.17 percent of properties repossessed to 0.8 percent. Though there has been a substantial fall since then, one can still assume that future mortgage failure rates will fluctuate considerably over time.

Figure 1
Building Society Possessions and Arrears (*Source: BSA & CML*)



Also, the cost to a bank of defaults on mortgage repayments varies according to the value of the property on which the mortgage is secured. This in turn depends on the change in housing prices since the mortgage was established. In the U.K. the number of mortgage failures was highest at the same time that the cost to the banks was highest, due to falling housing prices. Theoretically, the cost of default to the bank is a compound distribution formed of the probability distribution of defaults and the probability distribution of housing prices (or, more accurately, the difference between the mortgage plus arrears and the value of the house on forced sale). Suitable econometric models of either of these quantities have not been developed for the U.K.; we therefore use the empirical distribution for the cost of default from past data to estimate the sensitivity of the internal rate of return.

To examine the impact of changes in mortgage default rates and housing price inflation we calculate the internal rate of return from mortgage lending using historical data for default rates and house values. The loan model is the same as in Section 3 except that we ignore early repayments. The default rate (q_t) and the loan loss fraction (f_t) are determined from data. Specifically, the loan loss fraction is set by

$$f_t = \max\{0, 0.05 + \frac{H_t - L_t}{L_t}\} \tag{15}$$

where H_t is the housing price at time t. We use a national index of housing price inflation to determine H_t/H_0. The initial housing price is related to the initial size of the loan via the loan-to-value ratio. We consider the extreme case where the initial loan-to-value ratios for mortgages that end in possession are all 100 percent. This maximizes the loss (f_t) and the impact of mortgage defaults on the banks' profitability. The quantity 0.05 in equation (15) represents accumulated arrears and any markdown that occurs when a possessed property is sold.

Data are available on the probability of mortgage failure in a particular year. What are needed for our calculations, however, are conditional probabilities. For example, the probability that a mortgage issued in 1985 failed in 1990, the probability that a mortgage issued in 1986 failed in 1990, and so on.

Assumptions are needed to enable us to estimate the relevant probabilities. We introduce two functions, $\phi(x, y)$ and $Q(x, y)$, which are defined by

$$\begin{aligned} \phi(x,y) &= \text{Pr[Mortgage fails during month } y \mid \text{Mortgage started in} \\ &\qquad \text{month } x \text{ and survived to the start of month } y] \\ &= q_y \bar{N} Q(x,y) \end{aligned} \tag{16}$$

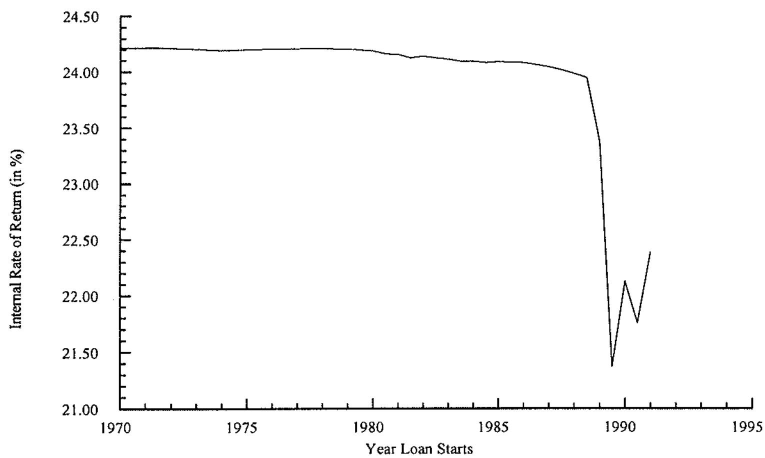
where q_y is the British national default rate (regardless of month of mortgage origin) in month y, according to British data; and $\bar{N}$ is the average length of a mortgage (taken to be 84 months); and $Q(x, y)$ is

$$Q(x,y) = \begin{cases} (y-x)/1200 & y-x = 1,2,\ldots,24 \\ 25/1200 & y-x = 25,26,\ldots,48 \\ (73-(y-x))/1200 & y-x = 49,50,\ldots,72 \\ 0 & y-x = 73,74,\ldots. \end{cases} \tag{17}$$

The *NPV* for a loan is calculated using equation (12) but with the default probabilities (q_t) replaced by $\phi(x, y)$ of equation (16) and the loan loss fraction given by equation (15).

Figure 2 shows the internal rate of return calculated by the cash flow model using data for default rates and housing price inflation. *IRR* is calculated for successive cohorts of loans. As we assume that no defaults happen more than six years after a loan is made, housing prices and default rates beyond 1997 will have no effect on the profitability of loans issued in 1991 or earlier. The results are calculated using a lending rate of interest $r_L = 10.5$ percent.

Figure 2
Variability of Returns Due to Changes in Default Rates and House Price Inflation



IRR is insensitive to default rates prior to 1989. Average housing prices peaked in the third quarter of 1989 and fell 12 percent over the next four years. The number of possessions peaked in the second half of 1991. Even in this severe time for the housing market the internal rate of return, based on the assumptions in our model, would have fallen only 3 percent. This illustrates the relatively low risk of mortgage lending due to loans being secured by the value of the borrowers' houses.

The model developed is flexible; for an unsecured loan the compound distribution for the cost of default will depend on default rates (which could be similar to those for mortgages) and the fraction of the loan recovered (which could be less than for mortgages). The variability of *IRR* probably would be much greater. Banks should take into account this risk of default both when setting interest rates that compensate the bank for default and when setting the hurdle rate of return, which should depend on the variability of returns. The hurdle rate should be higher for riskier (unsecured) loans.

5 The Pricing of Features

In this section the net present value will be calculated as a function of interest rate for loans with a variety of features. The purpose is to show the use of cash flow models in assessing how expensive such features are. The loans considered here are not identical to the ones used in the previous sections.

5.1 Cash Back

Many loans are provided that either give the customer some extra cash at the outset of the loan or offer some discount on the loan rate charged for the first year. These features are designed to attract customers to the bank. An alternative would be to offer a constant rate of interest throughout the loan that would be lower than the rate charged in the cash back scheme and lower than the rate charged beyond the first year in the discount scheme.

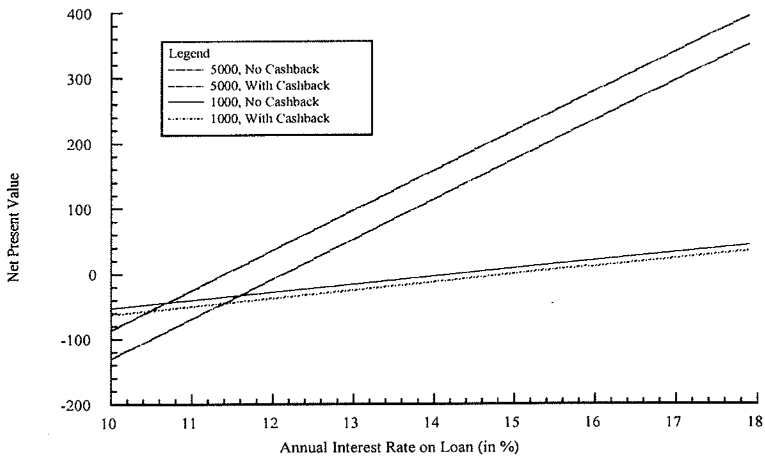
Cash back is included in the cash flow models by treating it as an extra initial expense. For example, if the money needed to provide the cash back is borrowed from the bank's treasury, it may be included in equation (14) by replacing E_0 with $E_0 + CB$ where CB is the amount of cash back. If cash back is paid out of capital, it may be included in equation (11) by replacing E_0 with $E_0 + CB$.

Figure 3 shows the effect that cash back of 1 percent of the loaned amount has on the net present value. The underlying parameters of these loans are the same as used in Example 2. (There are no early repayments or defaults in these calculations). The annual interest rate required to achieve a given NPV is higher by 0.72 percent for both loan sizes; i.e., this is the cost of the cash back. This is a substantial difference in a competitive loan market.

Figure 3 also illustrates a couple of other points: (i) the importance of loan size on the interest rate required to make lending sufficiently profitable (2.9 percent extra for the smaller loan here), and (ii) the greater sensitivity of the profit to interest rates for the larger loan. The slopes of the lines are roughly proportional to the loan size.

Figure 3 is produced using the proportional repayment method for funds and assuming that the initial expenses are paid for by borrowing

Figure 3
Effect of Providing Cashback of 1% of Loan



from the treasury. If initial expenses and cash back are both paid from capital rather than borrowing and the hurdle rate is unchanged, each *NPV* line in Figure 3 will shift to the right (i.e., a higher loan rate is needed to produce a given *NPV*). Small loans are affected more strongly than large loans because the initial expenses are proportionately larger, and likewise the loans with cash back have a larger increase in break-even loan rate than those without cash back (again, because setting aside the capital is expensive).

Table 11 shows the parameters that interact with cash back and those that do not. It also shows how much the break-even loan rate increases if cash back of 1 percent or 5 percent is provided. The values are given for two methods of paying for initial expenses (and also cash back, as this is treated as an additional initial expense), i.e., borrowing or using capital.

These values show that the amount of cash back is important to the break-even loan rate, with the change in this rate being five times greater for the 5 percent cash back loans than for the 1 percent cash back loans. The method of financing the cash back is also important, with the capital payments requiring a larger increase in loan interest rate for a given level of cash back than if the money is borrowed from the treasury.

Table 11
Change in Break-Even Loan Rates, Δr_L, in Percentage Points For Various Levels of Cash Back (CB)

Parameter Changed	New Value	Borrowing 1% CB	Borrowing 5% CB	Capital 1% CB	Capital 5% CB
Standard		0.75	3.75	0.85	4.24
L_0	3,000	0.75	3.75	0.84	4.24
	10,000	0.75	3.74	0.85	4.24
K_0/L_0	2.5%	0.75	3.76	0.84	4.23
	10%	0.75	3.75	0.85	4.26
n	18	1.44	7.27	1.54	7.79
	60	0.47	2.35	0.57	2.84
r_H	15%	0.76	3.81	0.81	4.06
	25%	0.74	3.69	0.88	4.42
E_0	0	0.74	3.74	0.84	4.23
	100	0.75	3.75	0.85	4.24
E_t monthly	1	0.75	3.75	0.85	4.24
q_t	0%	0.73	3.64	0.82	4.13
	0.4%	0.77	3.86	0.87	4.36
f_t	0.1	0.75	3.74	0.84	4.23
	0.5	0.76	3.76	0.85	4.26
	1	0.76	3.78	0.85	4.28
G_t	0%	0.75	3.75	0.85	4.24
	2%	0.75	3.75	0.85	4.24
R_t $(t > 12)$	0%	0.74	3.72	0.84	4.21
	0.4%	0.75	3.77	0.85	4.26

Of the ten parameters varied in Table 11 (and for the range of values examined), seven have negligible interactions with the cost of cash back: the size of the loan, the capital backing for the loan, the initial expenses, the running expenses, the loan recovery fraction, early repayment fees, and early repayment rates (but see the comment on default rates below).

By far the most important factor in terms of the cost of cash back (other than the amount of cash back) is the length of the loan. A short loan requires the same cost to be met by fewer monthly payments; hence, a greater interest rate margin is needed. Likewise, a lower margin is needed for longer loans.

The two other parameters that have some (albeit minor) impact on the cost of cash back are the loan default rate and the hurdle rate of interest. The default rate's influence is due to the alteration of the average duration of the loan. (The early repayment rate is less important because the model excludes any repayments in the first year, so the impact on average duration of a change in this rate is smaller than that of the default rate.) The influence of the hurdle rate is more important for the loans where the initial expenses and cash back are paid using capital. A low hurdle rate makes the initial capital outlay on cash back less expensive in terms of the size of future positive cash flows demanded and does not require such large margins to be paid by the borrower. The cost of cash back is increased at low hurdle rates when the money for it is borrowed (this cost is decreased if capital is used).

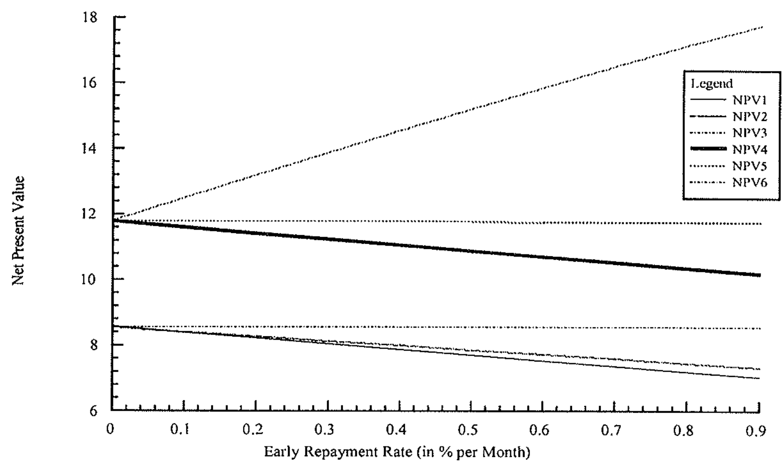
5.2 Early Repayment and Fees

For the loans examined, early repayments are not a significant factor in terms of loan pricing. For example, using the parameters in Example 3, which has no early repayments, the break-even loan rate is 11.98 percent. If early repayments happen at the high rate of 3 percent per month for the second and third years of the loan (in which case more than half of the loans are repaid early) and no early repayment fee is charged, the break-even interest rate only rises to 12.05 percent. Early repayments are more significant if the ratio of initial expenses to the size of the loan is high and the loan is short. Nevertheless, a bank may prefer to set the interest rate appropriate to the full term of the loan and charge early repayers a fee to compensate for missed future interest payments. We find in Allan et al., (1998) that early repayments are a greater problem if higher expenses are assumed at the outset.

The fees necessary to maintain *NPV* in the event of early repayment of loans have been calculated for two loan sizes (1,000 and 5,000) where the interest rates charged on the loans are 15 percent and 11.6 percent, respectively, and other parameters are the same as for previous examples except that there are no defaults and no cash back. Initial expenses are financed by borrowing from the treasury. There are no repayments in the first 12 months, and the rate is constant thereafter.

Results are shown in Figure 4 where *NPV* is plotted against the early repayment rate (R_t). Two fans of three lines are shown; the upper is for the $L_0 = 5,000$, $r_L = 11.6$ percent model, the lower is for the other pair of values. In each case the lowest line of the three lines is *NPV* if no fee is charged for early repayment. (The proportional fund repayment scheme has been used here so that the lines slope downward as early repayments increase.) Let G_t denote the fee actually charged for early repayment at time t. The middle line in Figure 4 reflects a fee of 0.6 percent of the loan outstanding, i.e., $G_t = 0.006L_t$, and the upper line reflects a fee of 2.9 percent of the loan outstanding, i.e., $G_t = 0.029L_t$.

Figure 4
Effect of Early Repayment and Fees

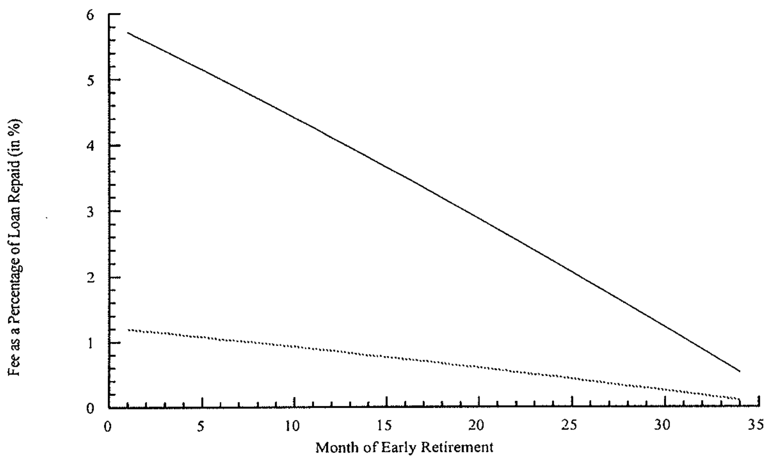


For the smaller loan the repayment fee needs to be at least 2.9 percent of the outstanding loan if the bank is not to lose by allowing repayments. But for the larger loan a proportionally smaller fee (i.e., about 0.6 percent) is needed. Thus, loan size is relevant to the impact of early repayment fees. The fee expressed as a percentage of the loan is inversely proportional to the loan size (although, the ratio 2.9 percent to 0.6 percent is close to the inverse of the ratio of loan sizes, i.e., 1,000 to 5,000).

Figure 4 shows that there is a linear relation between early repayment rate and net present value. It is possible to find a single fee (for a given set of loan parameters) that makes the bank broadly neutral to the frequency of early repayments.

The cost of an early repayment depends on the time that it happens, with the earlier repayments being more of a problem than those that happen close to the full term of the loan. Figure 5 shows the early repayment fee necessary to keep *NPV* of a loan constant; this fee changes with time and is calculated using a prospective approach. The fee is equated to the present value, at the time of early repayment, of the future net monthly income that would have accrued had there been no early repayment. There are also adjustments for the early return of capital less the repayment of initial expenses that had been borrowed from the treasury. The early repayment fee is given by

Figure 5
Early Repayment Fees



$$G_t = E_0 \frac{L_t}{L_0} - K_t + \frac{1}{P_t} \sum_{j=t+1}^{n} NMI_j (1 + i_H)^{-(j-t)} \tag{18}$$

where *NMI* is defined in equation (14) and does not include any adjustment for early repayments. (In the notation of equation (18), Figure 5 shows G_t/L_t vs. t.)

In Figure 5 the upper line is for the $L_0 = 1,000$, $r_L = 15$ percent combination and the lower line is for $L_0 = 5,000$ and $r_L = 11.6$ percent. Both lines are approximately, but not exactly, linear.

If the results are calculated instead for the situation where capital is used to pay the initial expenses (so the $E_0 L_t/L_0$ term is not needed in the above equation) and a higher interest rate is charged because this method is more expensive, the outcome is almost identical. This indicates that the costs of early repayment are not significantly dependent on the method used to pay for the initial expenses.

We conclude that there are three approaches one can take to model early repayments and assess the risk. Each of the three approaches can be used with the 25 year (full-term) loan model used here or with the seven year (average term) loan model used in Allan et al., (1998).

The first approach takes a best estimate of repayment rates and price to determine the correct average price charged to all borrowers. The problem with this approach is that it is deterministic and the risk of changed early repayment rates only can be assessed by deterministic scenario testing.

The second approach uses an actuarially neutral charging structure so that repayment fees can be charged for early repayments at any time. The fee would leave the internal rate of return of the loan unchanged whatever the time of surrender. Such an approach, which the authors believe will develop further in the U.K., would pass repayment risk to the borrower. It, therefore, would not require a stochastic approach.

The third approach involves a stochastic model of repayments to assess the variability of the internal rate of return given a reasonable model of early repayments. Repayments depend on the repayment fee and the degree of competition in the mortgage market. We believe that such an approach is unnecessary in the U.K. but may be necessary where the market is resistant to actuarially neutral early repayment fees. This is an area we leave for further research.

Our model has assumed a fixed interest rate throughout the term of the loan. In the U.K. most mortgages are variable rate. The results of the model would not be significantly altered if a variable rate were to be used as long as the interest margin (the difference between the loan rate paid by the borrower and the cost of funds for the bank) remained constant. It is the margin rather than the absolute level of interest rates that is important.

A more significant problem would arise when using the model to price fixed rate loans. It would be necessary to deal with the problem of borrowers exercising an option to repay early if variable interest rates fall. An option pricing approach to valuing that option could be used. Otherwise, three approaches are possible. First, the bank could investigate, using deterministic scenario testing, the effect of yield curve changes on profitability. The bank must make appropriate provisions for the exercise of an option to repay early. A second approach is to charge the borrower a penalty for early repayment. This is common in the U.K. The penalties are sufficiently high (for example, six months' interest for early repayment of a five year loan) to provide a significant disincentive for early repayment. Prepayment need not be a problem with the correct charging structure. Third, if the market will not bear prepayment penalties (or regulation prevents them), a fully stochastic model may be appropriate to enable the bank to assess the risk of prepayments caused by yield curve movements as well as other causes.

5.3 Risk Decreasing With Time

In this example the expected default rate and cost of default decrease with the time since the inception of the loan. The effect of varying the interest rate during the course of the loan to reflect this risk is examined.

For mortgages the risk to the bank is concentrated in the first few years of the loan. If a borrower has repayment problems later, the mortgage will be covered by the house value unless housing prices have fallen in nominal terms by more than the amount of the loan repaid. The bank could rearrange the mortgage or take possession of the property.

If risk decreases with time, shouldn't the interest rate? In practice the opposite is often the case as banks try to attract borrowers by providing low introductory rates.

The following example examines the possible benefit of varying the interest rate as the loan progresses. Several parameters are unchanged from previous examples: $r_H = 20$ percent, $r_F = 10$ percent, $r_C = 8$ percent, $r_{T2} = 10$ percent, and $E_t = 0$. Those parameters that differ from previous examples are $L_0 = 100,000$, $E_0 = 500$, $K_0/L_0 = {}^{T2}K_0/L_0 = 2.5$ percent, and $n = 25$ years. Also, default and repayment rates are different, and various values are considered for the loan interest rate. The defaults have the following pattern. The loan loss fraction is given by

$$f_t = \begin{cases} 0.05 + (L_t - L_{36})/L_{36} & t \le 36 \\ 0 & t > 36. \end{cases}$$

There are no losses after three years as the house value should be enough to cover the mortgage. Before three years the loss fraction decreases with time as part of the loan is repaid, but it is always at least 5 percent of the outstanding loan value. Otherwise, some form of rearrangement may be a more likely outcome than a default. In the first three years the default rate is either 0.2 percent per month or 0.4 percent per month. (Results are given for both values.)

Mortgage loans often end early as persons move before the repayment is complete or transfer to another bank before the full term is finished. The repayment pattern used is

$$R_t = \begin{cases} 0 & t \le 36 \\ 0.015 & t > 36. \end{cases}$$

It is not necessary that the early repayments start at the same time as the defaults stop, but a duration of one to five years is likely to be appropriate for both values. No fee is charged for early terminations of the mortgage in the model considered here. With this repayment pattern, roughly half of mortgages have ended by the seventh year.

How should interest rates be changed to take into account declining default risk? The equation for net monthly income for a set of loans (see, for example, equation (14)) includes

$$NMI = [i_L - (i_L + f)q]L + \text{Terms not involving defaults or } i_L.$$

This suggests that if $i_L - (i_L + f)q$ is kept constant as q and f change, the profitability of a loan will be broadly neutral to default risk. Let $\bar{q}$ and $\overline{(qf)}$ be defined as:

$$\begin{aligned} \bar{q} &= \frac{1}{n}\sum_{t=1}^{n} P_{t-1}q_t \\ \overline{(qf)} &= \frac{1}{n}\sum_{t=1}^{n} P_{t-1}q_t f_t. \end{aligned}$$

Then the invariant requirement leads to an interest rate set by

$$i_L^{Risk} = \frac{(1-\bar{q})i_L^{Mean} + (qf - \overline{(qf)})}{1-q}.$$

When the default rate in the first three years is 0.2 percent per month the weighted averages $\bar{q}$ and $\overline{(qf)}$ are 0.0727 percent and 0.0048 percent, respectively. When the default rate is 0.4 percent these two values are 0.1489 percent and 0.0098 percent, respectively.

The following results are obtained using a simpler formula for the risk interest rate, viz.:

$$i_L^{Risk} = i_L + \frac{q \times f - \overline{(qf)}}{1-q}$$

with i_L the same as the interest rate used in the fixed rate case. This equation is used because the relationship between the mean rate of interest charged under this regime to the rate charged in the fixed rate case is more readily apparent (i.e., they are the same).

With a varying interest rate the loan payment has to be recalculated each month. For month t, let $i_L(t)$ denote the interest rate in effect during month t, i.e., from time $t-1$ to t. Then the payments made at time t, X_t, and L_t are given by

$$\begin{aligned} X_t &= \frac{L_{t-1}}{a_{\overline{n+1-t}|\, i_L(t)}}, \quad \text{and} \\ L_t &= (1+i_L(t))L_{t-1} - X_t. \end{aligned}$$

The amount received by the bank is $(1-q_t)P_{t-1}X_t$.

Table 12 shows *NPV* for two default rates and four interest rates that are the smallest, to the nearest 0.01 percent, that give a positive *NPV* at these two default rates, for both the fixed interest rate and declining interest rate case. Two *NPV* values are given for each q and r_L combination, corresponding to these two rate setting methods. In the table r_L means both the rate used in the single rate calculation and the annual equivalent of the rate appearing in the above equation for the risk rate. The mean refers to the straightforward average of the interest rate over the term of the mortgage (not weighted by the survival probability; the weighted average is just r_L).

Table 12

***NPV* for Various q and r_L Combinations**

	Interest Rate				*NPV*	
q	r_L	Mean	Max	Min	F-Rate	V-Rate
0.2%	10.58%	10.54%	10.73%	10.52%	31.7	180.8
0.4%	10.58%	10.49%	10.87%	10.45%	-388.8	-105.4
0.2%	10.71%	10.67%	10.86%	10.65%	455.7	604.1
0.4%	10.71%	10.62%	11.00%	10.58%	16.9	299.0
0.2%	10.53%	10.49%	10.68%	10.47%	-131.3	18.1
0.4%	10.62%	10.53%	10.91%	10.49%	-263.9	19.0

Notes: F-Rate = Fixed interest rate; V-Rate = Variable interest rate.

The net present values are about 150 higher for the variable rate model in the $q = 0.2$ percent cases than in the fixed rate model and

280 higher for the $q = 0.4$ percent cases (for comparison, the initial capital outlay is 2500). Alternatively, instead of making a larger profit at the same (weighted average) interest rate, a lower average rate can be charged, as shown in the last two lines.

Moreover, the importance of the default rate is reduced when this flexible interest rate is used. For example if the rate had been set at 10.58 percent in anticipation of the lower default rate, the reduction in *NPV* that happens if the default rate proves to be 0.4 percent per month is 420 if the fixed rate is used and 286 if the flexible rate is used. This method only works if the risk adjustments to the interest rates are based on the 0.4 percent value. If they are not, the loss is just as great as in the fixed rate case.

This example illustrates that average interest rates can be reduced if risk-related pricing is introduced. The suggested differences in the average rates charged are not large, with reduction in the rates less than 0.1 percent.

6 Summary and Conclusions

Cash flow models that are used in other areas of actuarial work could be used in banking to price loans. Often bank lending is priced by looking at the whole book of business together and ensuring that the margin is adequate to provide the required return on capital on the business as a whole. Adjustments to the margin will generally be made to allow for the risk profiles of different borrowers. A cash flow approach would consider explicitly the cash flows that were expected for a particular category of loans. It would therefore be possible to set an interest margin, appropriate to that category that allowed explicitly for: the capital used to back a particular category of loans; the expenses of a particular category of loans; and the risk of a particular category of loans. The cash flow approach can also handle the many interest rates that are relevant to bank lending. The following interest rates give rise to cash flows: interest charged on the loan; interest paid by the lender to the source of the money lent; hurdle rate of return on equity capital; rate of interest to be paid on debt capital; interest earned on set-aside equity capital.

Having determined the cash flows pertaining to a particular category of loan, the cash flow model can be used to make business decisions such as determining the interest margin to be charged or determining whether the interest margin available on a category of lending in the market makes the loan sufficiently profitable. For marketing reasons, it may be appropriate to give cross subsidies between categories of loans. The cash flow model allows the effect of these cross subsidies to be quantified.

When lending is secured (for example mortgage lending) the effect of defaults is not significant but in general defaults are important. The model has been extended to deal with defaults but more empirical work is necessary to find reasonable models of default rates and the loss incurred by a bank on defaults. Various ways of dealing with prepayments are discussed. Prepayments can cause difficulty for two reasons: first initial expenses may not be recouped if a loan is prepaid; second, if a loan is given at a fixed rate of interest there is a financial interest rate option against the lender. In the U.K. market, it may be possible to develop charging structures so that early repayments do not have a material financial effect on the bank (prepayment penalties are common). Where this is not possible, stochastic modeling of prepayments should be performed.

The model is extended to include pricing for loan products which involve cash backs and prepayment fees. The sensitivity of the model to various parameters is tested and it is found that expenses; the interest margin; the size of loan; and the duration of the loan are the most important parameters. Banks may wish to differentially price loans to a greater extent than is currently the case to allow for size and duration of a loan. Alternatively, as has been mentioned above, prepayment fees or other charging structures could be used to ensure that loans which are prepaid are still profitable.

The main areas for further work are the development of better models for estimating the costs of default and the modeling of prepayments where charging structures do not make the bank indifferent to prepayments.

References

Allan, J.N., Booth, P.M., Verrall, R.J., and Walsh, D.E.P. "The Management of Risks in Banking." Presented on February 23, 1998 to the Institute of Actuaries. Forthcoming in *Institute of Actuaries Session Meeting Paper*, (1998).

Altman, E.I. *Corporate Financial Distress.* New York, N.Y.: John Wiley & Sons, 1983.

Altman, E.I. "Corporate Bond and Commercial Loan Portfolio Analysis." Wharton School Working Paper Series, pp. 96-41. Philadelphia, Pa.: University of Pennsylvania, 1996.

Brealey, R.A., and Myers, S.C. *Principles of Corporate Finance.* New York, N.Y.: McGraw Hill, 1991.

British Bankers' Association. *Annual Abstract of Banking Statistics. Banking Act Report.* London, England: Bank of England, 1995.

Building Societies Association. "Trends in Mortgage Arrears." *BSA Bulletin* 43 (1985): 18.

Davis, E.P. "Bank Credit Risk." *Bank of England Working Paper Series* 8 (1993).

Griffin, K. "The Retail Banking Industry in Australia." Institute of Actuaries of Australia session meeting paper (1996).

Hare, D.J.P., and McCutcheon, J.J. *An Introduction to Profit Testing.* London, England: Institute of Actuaries, 1991.

Higson, C.J. *Business Finance.* London, England: Butterworths, 1986.

Council of Mortgage Lenders. "Table 25" *Housing Finance* 32 (1996): 46.

Hylands, J.F., and Gray, L.J. *Product Pricing in Life Assurance.* London, England: Institute of Actuaries, 1989.

Lewin, G.C., Carne, S.A., DeRivaz, N.F.C., Hall, R.E.G., McKelvey, K.J., and Wilkie, M.A. "Capital Projects." *British Actuarial Journal,* 1 part 2 (1995): 155-249.

Squires, R.J. *Unit Linked Business.* London, United Kingdom: Institute and Faculty of Actuaries, 1986.

Taffler, R.J. "Forecasting Company Failures in the UK Using Discriminant Analysis and Financial Ratio Data." *Journal of the Royal Statistical Society,* Series A, 145 part 3 (1982): 342-358.

Appendix A: How Credit Risk is Assessed

Corporate Loans

A bank can group large borrowers according to expected default risk. The allocation of a borrower to a risk group generally is based on accounting ratios. Statistical techniques (regression or multiple discriminant analysis) are applied to historical data on bankruptcies or loan defaults with these accounting ratios to produce a set of weights for the ratios. Any potential borrower's accounts can be studied to provide a score that is intended to be a predictor of default risk.

The procedure has been in widespread use since the 1960s, and some of the formulae used have been published. It is often found, however, that weighting factors determined from one data set are not the same as those from another set: e.g., U.S.-derived weights are not applicable in the U.K., and 1970s values are not useful now.

A review of many analyses of corporate default is presented in Altman (1983). For a paper relating to risk factors in the U.K., see Taffler (1982). References to more recent papers are found in Altman (1996). Altman (1996) also provides a formula for a score that is a predictor of default risk, with a higher score indicating a lower probability of default:

$$Score = c_0 + c_1X_1 + c_2X_2 + c_3X_3 + c_4X_4$$

where

$$\begin{aligned} X_1 &= \text{Working capital/total assets;} \\ X_2 &= \text{Retained earnings/total assets;} \\ X_3 &= \text{Earnings before interest and taxes/total assets; and} \\ X_4 &= \text{Equity (book value)/total liabilities.} \end{aligned}$$

Also, c_1, c_2, c_3, and c_4 are positive constants.

Personal Loans: Credit Scoring

The method of assessing the risks in personal loans is called *credit scoring*. The statistical techniques used are similar to those used for large corporate loans (e.g., discriminant analysis), but the factors in the model change (there are no audited accounts to use).

There is considerable danger in using past data to predict future bad debts. The economic background is likely to influence the overall level of bad debts. Within this overall trend, the credit score should indicate some ranking of risk.

For a given type of loan, the most important data are provided by credit agencies. Evidence of existing bad debts with other banks and evidence of successful maintenance of credit repayments are relevant. Other demographic data as provided on loan application forms may be used, but are not always major influences. (One reason is that banks find that short application forms are useful in attracting customers.) A third source of information is behavior, i.e., the credit history of a loan applicant who is an existing customer of the organization. All of the factors can be combined onto a scorecard, and historical data will provide default rates and costs of default versus score.

The construction of a scorecard usually is done by a specialist agency. Typically a new scorecard will be prepared every two or three years. Around 1,000 to 2,000 bad loans are needed (and an equal number of good loans) to provide statistically sound weights. Given that a default rate of only a few percent a year is not uncommon, this number of bad loans requires a large portfolio, a long base period, or a weak definition of bad. In practice the last option is likely to be selected.

It is appropriate to have different scorecards for different products (e.g., mortgages and credit cards) and for different categories of customer (e.g., new or existing customer).

There is an important limitation on the reliability of default rate predictions; the data are for a select set of the population, i.e., persons accepted for loans a few years ago. Risk factors that are unimportant among this group may be significant in the population of future loan applicants.

Once a score has been calculated for a loan applicant, the most common approach in bank lending is a straight accept or reject decision (e.g., accept the application if the score is greater than, say, 100). This contrasts with risk-based pricing in insurance where premium rates vary with risk rather than being one rate for all accepted. The equivalent response in terms of lending would be to charge a rate of interest that varies with risk. This method has been introduced in some areas of bank lending. An alternative method, used in practice with credit cards, is for a bank to operate several cards (perhaps under different names) with different interest rates. To be accepted for a low interest rate card, a higher score will be needed than for the higher interest rate cards.

Appendix B: Uniform Repayments to the Treasury

This section expands possible ways in which the bank could repay money to the bank's treasury.

The paper has been based on the assumption that the bank always owes to the treasury an amount equal to the amount that the borrower owes. (Initial expenses are also paid in the same pattern.) An alternative is that the initial amount borrowed from the treasury (for loans and initial expenses) are amortized over the period of the loans and paid in equal installments irrespective of whether some loans default or are repaid early.

If there is a difference between the amounts owed by the bank (to the treasury) and to the bank (by the borrower) this does not mean that there is idle cash available. The bank holds no money other than capital (which is invested in the wholesale cash market); all of the money it receives is immediately paid to the treasury or assigned to the providers of capital as profit.

Under the uniform repayment scheme the equation for net monthly income, which is comparable to the equation (14), is:

$$\begin{aligned} NMI_t &= XP_t^{(qr)}(1-q_t) - \frac{L_0+E_0}{a_{\overline{n}|\,i_F}} \\ &\quad + i_C P_{t-1}^{(qr)} K_{t-1} + (P_{t-1}^{(qr)} K_{t-1} - P_t^{(qr)} K_t) - P_{t-1}^{(qr)D} K_{t-1}(i_D - i_C) \\ &\quad - E_t + q_t P_{t-1}^{(qr)}(1-f_t) L_{t-1} + R_t(1-q_t) P_{t-1}^{(qr)}(L_t + G_t). \end{aligned}$$

(Everything is the same except for payments to the treasury.)

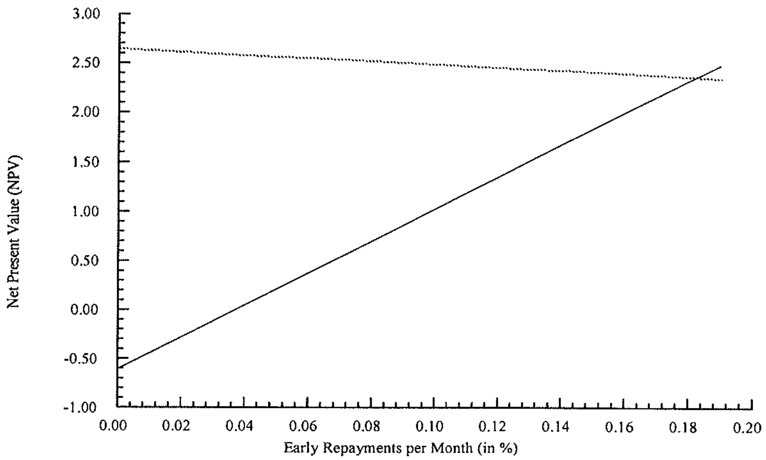
Figure B1 illustrates the significance of the two fund repayment patterns. The parameters used in producing this figure are the same as for the second spreadsheet example, except that early repayments are included (but there are no early repayments allowed in the first year). A loan interest rate of $r_L = 11.45$ percent is chosen to make *NPV* close to zero. (At the maximum early closure rate shown, 0.2 percent, 95 percent of loans last the duration.)

Figure B1 shows that even without premature repayments there is a difference in the value of this loan under the two methods of funding. The proportional method gives a higher value because there is a slight delay in the timing of repayments to the treasury. (These repayments are initially less than in the uniform case, but are greater later. In total they are a little larger under the proportional method.)

Also, early repayments reduce the value of the loan, as far as the bank is concerned, when the proportional method of fund repayment is used. Early repayments are welcome under the uniform scheme, however. Although it may seem unreasonable for a bank to want a profitable loan to stop, the improvement shown in the figure is not wrong. The result depends on the bank being able to secure funds from the treasury at a rate of interest below the hurdle rate that can be retained after some of the loans have been repaid (or defaulted).

The figure illustrates the point that the way the funding is arranged or the method of bookkeeping selected can be more important in de-

Figure B1
Different Repayment Patterns for Funds



termining the profitability of a class of loans than a key feature of the loan (in this case the early repayment pattern).

The steepness of the line for the uniform repayment case shows that the profitability of the loan is likely to be more volatile if this scheme is used. Also, the internal rate of return of loans financed by the uniform repayment scheme is more sensitive to the default rates, when the default rates are high than they would be under the proportional repayment scheme. In this sense there is more risk accepted by the providers of capital when a uniform repayment plan is used, and a higher return therefore may be required.

Stability of Representative Crediting Rate Scenarios Under Monte Carlo Simulations

Sarah L.M. Christiansen* and Kelley Buchacker†

Abstract‡

We develop a methodology to ensure that a Monte Carlo simulation of the distribution of the primary rates, used for determining an interest crediting rate, is stable regardless of the initial random number seed. We consider the implications of the use of antithetic random normal deviates upon the scenario process and modifications to the candidate list and the choice function within the representative process. It is shown that the use of antithetic random deviates alone does not have a statistically significant effect on our results. The other two modifications (candidate selection algorithm and choice function) are statistically significant. Furthermore, the synergistic effects of the antithetic random deviates, candidate selection algorithm, and choice function are significant.

Key words and phrases: *choice function, extreme selection, ANOVA, antithetic random number, fit measure, multinomial distribution*

*Sarah L.M. Christiansen, Ph.D., F.S.A., M.A.A.A., is an assistant corporate actuary at the Principal Financial Group. She received her B.A. in mathematics from the University of California-Riverside in 1966 and her M.A. (1968) and Ph.D. (1971) from the University of New Mexico. Before starting her actuarial career in 1985, she taught mathematics at Drake University, Des Moines Area Community College, and Grandview College.

Dr. Christiansen is currently a member of the following Society of Actuaries' committees: Finance Research (Chair); Finance Practice Advancement; Research Project Oversight; Finance and Investment Education Objectives; the Spring Program; and the Education and Research Section Council.

Dr. Christiansen's address is: The Principal Financial Group, 711 High Street, Des Moines IA 50392-0650, U.S.A. Internet address: *Christiansen.Sarah@Principal.com*

†Kelley Buchacker is an assistant actuarial statistical analyst at the Principal Financial Group. She received her B.Sc in statistics from Iowa State University in 1994.

Ms. Buchacker's address is: The Principal Financial Group, 711 High Street, Des Moines IA 50392-0650, U.S.A. Internet address: *Buchacker.Kelley@Principal.com*

‡The authors would like to thank the Principal Financial Group for its support and the anonymous reviewers for their helpful suggestions.

1 The Problem

A client of the Principal Financial Group's corporate actuarial department is interested in using randomly generated interest-rate scenarios[1] for determining its interest-crediting rate. Crediting rates are revised monthly or whenever a major shift in interest rates occurs, but there are no plans for daily rate determination. For this reason, a methodology that is more robust than the usual arbitrage-free methods is preferred.

Prior to generating scenarios, a random number generator is chosen. The random number generator can be one created especially for the simulation at hand or the one included with the programming language or software.[2]

A single scenario consists of the original yield curve followed by the 30 years of simulated yield curves and is generated using Monte Carlo techniques as follows:

1. The original yield curve is determined based on the date of the scenarios;

2. Next, a random number is chosen to be the seed, that is the starting value, for the random number generator;

3. Two sequences of 30 random numbers are obtained from the generator: one sequence is used to determine the long rates and the other sequence is used to determine the shape codes;

4. Finally, these random numbers and the original yield curve are used to generate a sequence of 30 years of yield curves called the interest rate scenario.

The client, however, first must be certain that similar crediting rates would result from similar yield curves. In testing the original method of determining representative interest rate scenarios, the client found that similar curves were leading to crediting rates that were not sufficiently close. The client's concern is that the random number seed, which is selected in a random manner, may have a significant impact on the results of the pricing runs. Using a deterministic set of seeds for the random number generator will result in the scenarios changing from stochastic

[1]An interest rate scenario consists of a set of yield curves. There is a curve for each of the next 30 years (at least one per year). A rate on each curve is specified for each of the following maturities: three months, six months, one year, two years, three years, four years, five years, seven years, ten years, 15 years, 20 years, and 30 years.

[2]For more on the construction of random number generators see, for example, Kalos and Whitlock (1986, Appendix) or Bratley, Fox, and Schrage (1983, Chapter 6).

to deterministic (albeit in a fashion that is not obvious). The challenge is to modify the method for generating representative scenarios so that the resulting scenarios exhibit sufficient stability.

2 Representative Scenarios

2.1 Definition of Representative Scenarios

For a given set of scenarios, a representative scenario is a subset of the set of scenarios that has, across all maturities simultaneously, approximately the same mean, median, range, and variance as the entire set of scenarios. There are 13 maturities considered in this process: three month, six month, one year, two year, three year, four year, five year, seven year, ten year, 15 year, 20 year, and 30 year rates and the shape code. Although the shape code is not a maturity, it is the only variable that relates the values between maturities to each other. For that reason, we include it in our discussion as if it were a maturity. For convenience, we label the maturities in Table 1.

Table 1
Definitions of Labels

1	Shape code
2	Three months
3	Six months
4	One year
5	Two years
6	Three years
7	Four years
8	Five years
9	Seven years
10	Ten years
11	15 years
12	20 years
13	30 years

The original method for determining the representative scenarios is described in Christiansen (1998). The representative scenario process involves the following steps:

1. Generate a set of 1,000 scenarios;
2. Split these scenarios into five subsets of 200 scenarios each;
3. Choose an algorithm to determine potentially representative scenarios;
4. Use the algorithm to reduce each subset of 200 scenarios to a subset of ten representative scenarios called the *best candidate*.[3]

This process results five best candidates giving a total of 50 representative scenarios.[4]

2.2 Why 1000 Scenarios?

The rationale for testing sets of 1,000 scenarios is the ability to distinguish differences due to the representative processes from those due to the underlying data. This is another way of asking whether the original sample of size 1000 is adequate. Two methods for determining the adequacy of sample size are suggested in the literature. One method due to Greg Taylor was introduced at the 1994 Risk Theory Conference (Oberwolfach, Germany). Taylor's method looks to see when the sample variance converges to the variance of the underlying distribution.

Robbins, Cox, and Phillips (1997) suggest a variation of the Taylor method: plotting the variances using progressively larger samples to determine the number of scenarios required for the variance to converge. In their example, convergence begins at a sample size of around 400. While the number of scenarios required for the variance to converge depends on the application, their examples suggest that samples of size 1,000 may be adequate (i.e., exhibit a reasonable amount of stability).

[3]A *candidate* is a subset of ten scenarios that are potentially representative scenarios. A *candidate list* is the set of all candidates selected by the algorithm being used.

[4]Here we are defining best candidates and representative scenarios with respect to the subset of 200 scenarios to which they belong.

3 Improving Stability

Three possible modifications to improve the stability of results were identified in our discussions with the client: (i) modify the way random variables are generated by using antithetic normal random variables to generate scenarios; (ii) modify the selection criteria; or (iii) modify the choice function.

- Antithetic Variates: The use of antithetic normal random deviates is a well-known variance reduction technique for Monte Carlo simulations. The technique exploits the decrease in variance that occurs when random variables are negatively correlated. The hope is that when variables are negatively correlated, a random variate x yields a value above (below) the mean, then $(1 - x)$ is likely to be below (above) the mean, and the average is likely to be closer to the mean. To generate a set of antithetic normal random variables one half of the set is generated and their negatives are used for the second half. See, for example, Kalos and Whitlock (1986, Chapter 4.4) and Tilley (1987). Ideally, their use would not have any impact on the overall scenarios, while helping add stability to the representative scenarios. Therefore, no statistically significant differences were expected due to the use of the antithetic random normal deviates.

- Selection Criteria: Use an extreme selection criteria to determine the candidate lists rather than the alternative two standard deviation selection method. An extreme is considered because it generates a much larger, but still manageable, list of candidates from which to choose. The nature of matched extremes, however, may be misunderstood. A match between the minimum or maximum in a single scenario and maturity and the overall minimum or maximum does not mean that the average of the rates for that maturity and scenario is the minimum or maximum. We expect that the larger candidate list will make a difference.

- Choice Function: Revise the choice function that is used to select the candidates from the candidate lists. The purpose of the choice function is to select the most desirable candidate from the candidate list. The choice function is a weighted sum, over all maturities, of the absolute deviations between the means of the candidate and the corresponding means of the set of 200 scenarios. The best candidate is the one whose choice function is the minimum.

These three possible modifications are independent and may work synergistically. Therefore we consider all possible combinations of these three modifications. Each combination is referred to as a *method* (Table 2).

Table 2
Definitions of Methods

	Choice Function	Candidate List	Random Numbers
Method 1	New	Extremes	Base
Method 2	New	Extremes	Antithetic
Method 3	New	2 Std Dev	Base
Method 4	New	2 Std Dev	Antithetic
Method 5	Old	Extremes	Base
Method 6	Old	Extremes	Antithetic
Method 7	Old	2 Std Dev	Base
Method 8	Old	2 Std Dev	Antithetic

Notes: Base refers to non-antithetic random variables. Thus, if you want 100 base random numbers, for example, you simply generate 100 independent random numbers from the generator.

4 Creating Candidate Lists

We test two methods for selecting the candidate list: the matched extremes method and the two standard deviations method. *Matched extremes* are defined as those in which the selected scenario has either the same minimum or maximum rate at some point in time as does the subset of 200 scenarios for that maturity. Both of these methods are based on maturity (excluding shape code) so $i = 2, 3, \ldots, 13$.

4.1 Matched Extremes Method

This method of creating candidate lists begins by creating a list of all possible combinations of matched extremes for the maturity under consideration, say maturity i.

For maturity i, all combinations of the matched extreme scenarios, one matching the minimum rate, the other matching the maximum rate, without repetition, form the first two elements of the candidates' (in maturity i) contribution to the candidate list.[5] For maturity i, the mean (μ_i) and standard deviation (σ_i) are determined for the subset of 200 scenarios. Also the mean (m_i) for each scenario is computed and the scenarios (whose mean is closest to each of these values $\mu_i - 0.85\sigma_i$, $\mu_i - 0.65\sigma_i$, $\mu_i + 0.65\sigma_i$, $\mu + 0.85\sigma_i$), are added to each combination of candidates in the candidate list (while avoiding any duplicate scenario numbers).

Each candidate now consists of six (of the required ten) scenarios; the remaining four scenarios selected are the four scenarios whose means are closest to $(10\mu_i - 6m_i)/4$. This choice for the final four scenarios ensures that the average of the candidates is as close as possible to μ_i for the maturity under consideration. This candidate list varies in length, but generally consists of approximately 200 to 500 candidates.

4.2 Two Standard Deviations

For maturity i, an alternative method is to de-emphasize the extremes used in the representative scenario process. The candidate list consists of two candidates per maturity. The first candidate replaces the minimum from the matched extremes method with the scenario whose mean is closest to $\mu_i - 2\sigma_i$ and replaces the maximum with the scenario whose average is closest to $\mu_i + 2\sigma_i$. The remaining eight scenarios for this candidate are selected in the same manner as the matched extremes method.

To select the second candidate the 200 scenario means are arranged in increasing order. The first six scenarios then are chosen by arranging the means in increasing order and selecting the 10th, 20th, 30th, 170th, 180th, and 190th, respectively. The last four scenarios for the second candidate are chosen such that the overall mean is equal to the mean of the subset of 200 scenarios. Thus, this alternative method produces a candidate list with 24 choices (two for each of the 12 maturities).

5 The Choice Function

5.1 The Old Choice Function

The mathematical form of the old choice function, $C^{(\mathrm{old})}(k)$, is

[5] See Section 6 for an example on how this is done.

$$C^{(\text{old})}(k) = \sum_{i=1}^{13} w_i \times |\mu_i - m_{i,k}| \tag{1}$$

where k denotes the candidate, i denotes maturity, $m_{i,k}$ denotes the mean for candidate k and maturity i, and the w_i's are non-negative weights related to the importance of the maturity in terms of the purpose for which the scenarios are to be used. Note that the averages are taken over all of the times in the future (i.e., over the next 30 years) and all of the scenarios in the candidate for $m_{i,k}$ and all of the scenarios for μ_i.

To be more specific, let $r_{n,t,i}$ be the interest rate associated with scenario n, at time t, and maturity i. We can then define the following

$$\begin{aligned} m_i^n &= \frac{1}{31} \sum_{t=0}^{30} r_{n,t,i} \\ \mu_i &= \frac{1}{200} \sum_{n=1}^{200} m_i^n \\ m_{i,k} &= 0.1 \sum_{n \in \text{Candidate } k} m_i^n. \end{aligned}$$

Note that both μ_i and $m_{i,k}$ are averages of the average rate (by time) for some scenarios. In the case of μ_i, all 200 of the scenarios are included in the average, whereas for $m_{i,k}$, only those scenarios that are included in candidate k are in the average.

A potential shortcoming of the old choice function is that it does not differentiate between different scenarios with the same average rates. For example, consider two different scenarios: one scenario has rates that first declined and then rose, and the other scenario has rates that first rose and then declined. Given a situation where the majority of the liabilities would be gone by the end of the first ten years, then each scenario would have a different impact on the results of the crediting rate strategy.

The client was concerned that the matched extremes method (5) would over-emphasize the extremes of the scenarios and had, therefore, selected the two standard deviation method (7). When the client tested scenarios created by method 7 (in their model), however, they found differences of 40 basis points could arise from similar yield curves (run at different times). Testing the scenarios that encompass a 30 year period shows that the 40 basis point difference in the proposed crediting rate could be attributed to differences in the means of the first ten years of the seven year rate. These differences are the impetus for our study.

5.2 The New Choice Function

Modifying the algorithm for selecting the representative scenarios to acknowledge the timing is a simple task. The choice function $C^{(\text{new})}(k)$ is revised to reflect the timing concerns as follows:

$$C^{(\text{new})}(k) = \sum_{i=1}^{13} w_i \times \left[|\mu_i^A - m_{i,k}^A| + f \times |\mu_i^F - m_{i,k}^F| \right], \tag{2}$$

where the superscript A indicates that the mean is taken over all 30 years in the scenario, while the superscript F indicates that the mean is taken over first ten years. The constant factor f indicates the relative importance of the first years to all of the years. In our study, the weights for the first ten year means for all maturities were arbitrarily set to twice the weights for the same maturity for all years, i.e., $f = 2$. This new function looks like the old choice function, but has twice as many terms.

The reasons for considering this particular revision of the choice function is related to the business being modeled. Here, most of the liability cash flows are gone by the end of the ten year period. Therefore, it does not make sense to treat all of the years of the scenarios equally. We anticipate that this change will be significant and beneficial.

To gain the maximum benefit from revising the choice function, the same concerns should be reflected in the choice of the candidate list. Therefore, the candidate list is revised and selected based on two separate criteria: (i) the first ten years of the scenario (times 0–10), and (ii) the entire time horizon (times 0–30). These criteria do not impact the manner in which the candidate list is selected (either matched extremes or the two standard deviation method). In the case of the two standard deviations method, however, there now are 48 candidates instead of 24.

6 A Candidate List and Choice Function Example

This example is an excerpt from a single run of 200 interest rate scenarios and is based on the matched extremes method. The data used to create the candidate list were taken from a computer run. Table 3 contains data from the three-year maturity for 25 of the 200 scenarios.

From Table 3, we see that scenarios 70, 74, 82, 108, 111, 116, and 186 have a minimum value of 3.5, the same as the minimum value for the run of 200 scenarios. Scenario 40 is the only case that matches the maximum value. The contribution to the candidate list from the three-year rate, begins as shown in Table 4.

If, in addition, scenario 82 matched the maximum, then we would have six additional lines with 82 replacing 40 (i.e., 70 82; 74 82; 108 82; 111 82; 116 82; and 186 82).

Table 3
Three-Year Maturity Data for 25 of 200 Scenarios

Scenario	Average	Minimum	Maximum
1	6.516	4.261	9.046
2	6.464	4.283	9.252
40	10.379	6.620	20.138
46	9.335	5.856	12.944
52	6.252	3.854	11.238
54	7.237	5.167	10.322
65	7.324	4.241	12.303
70	5.774	3.500	8.641
74	5.017	3.500	8.065
75	9.019	5.436	13.077
82	4.690	3.500	8.264
86	7.140	4.211	11.694
107	9.181	6.066	13.567
108	6.322	3.500	10.439
110	8.177	6.050	10.967
111	7.253	3.500	13.898
112	7.018	4.504	9.895
115	7.452	5.027	10.487
116	5.844	3.500	9.399
121	7.222	4.334	11.264
124	6.256	4.681	8.052
146	5.894	3.507	9.596
169	8.701	4.930	13.752
180	7.114	4.639	10.113
186	5.068	3.500	8.165
All 200	7.428	3.500	20.138
All 200 Standard Deviation = 2.261			

Table 4
First Two Entries of Sample Candidate List

1st	2nd
70	40
74	40
82	40
108	40
111	40
116	40
186	40

Next from the data on all 200 scenarios at the bottom of Table 3, we calculate $\mu \pm 0.85\sigma = 7.428 \pm 0.85 \times 2.261 = 9.35$ or 5.506. We then look for the two scenarios whose average values are closest to 9.35 or 5.506. In reality, the best choices are not among the 25 scenarios illustrated. From the illustrated scenarios, we would select scenarios 46 and 70; although for the first candidate we must use scenario 116 instead of scenario 70 to avoid duplication.

Table 5
First Six Entries of Sample Candidate List

1st	2nd	3rd	4th	5th	6th
70	40	46	116	75	146
74	40	46	70	75	146
82	40	46	70	75	146
108	40	46	70	75	146
111	40	46	70	75	146
116	40	46	70	75	146
186	40	46	70	75	146

Next we find $\mu \pm 0.65\sigma = 7.428 \pm 0.65 \times 2.261 = 8.898$ or 5.958, and select the two scenarios closest to 8.898 or 5.958. The closest scenarios in the list of 25 are scenarios 75 and 146. Our candidate list now has six scenarios per candidate as illustrated in Table 5. Finally to fill out the candidate list, we perform the calculation for the first row only:

$$\begin{aligned} 10\mu - \sum m_{3\text{yr}}^{\text{sn}} &= 74.28 - 5.774 - 10.379 \\ &\quad - 9.332 - 5.844 - 9.019 - 5.894 \\ &= 28.138 \end{aligned}$$

Dividing this result by 4, we would like the remaining four scenarios to have an average three-year rate as close as possible to 7.034. The values can be below as well as above. In this case all four scenarios have a greater rate: 112, 180, 86, and 121.

Once the candidate list has been completed, the averages of all of the candidates for all of the maturities and the shape code are calculated, and the choice function is calculated. We present a small sample from the candidate list and choice function for the run.

Table 6
Sample of Candidate List

Label	Candidate List								
(1):	146	71	13	4	156	27	184	86	69
(2):	186	71	13	4	156	27	7	104	19
(3):	52	40	59	4	146	148	157	132	113
(4):	70	40	13	46	36	134	180	143	58
(5):	74	40	13	46	36	134	12	79	182
(6):	74	40	70	68	189	197	54	112	58
(7):	186	40	70	68	189	197	112	111	54
(8):	70	40	146	46	124	169	30	171	77
(9):	82	40	146	46	124	169	86	54	112
(10):	186	40	146	46	124	169	121	111	180
(11):	74	127	116	75	108	107	150	149	143

The candidate list in this example has approximately 500 entries. The result from the set of all two hundred scenarios is the goal. We would like to match all of these data simultaneously. Table 6 shows the candidate list. Table 7 shows the various statistics needed by the choice function to determine the representative scenarios.

Table 7

Candidate Statistics

Label	Shape	3 mos	6 mos	1 yr	2 yr	3 yr	4 yr	5 yr	7 yr	10 yr	15 yr	20 yr	30 yr	Choice
(1):	4.352	6.279	6.361	6.659	7.111	7.410	7.620	7.784	7.961	8.093	8.236	8.461	8.644	1.342
(2):	4.529	6.274	6.345	6.642	7.072	7.329	7.507	7.634	7.793	7.914	8.055	8.242	8.403	5.049
(3):	4.487	6.183	6.286	6.621	7.135	7.457	7.660	7.806	7.977	8.100	8.254	8.464	8.628	0.754
(4):	4.142	6.099	6.219	6.600	7.128	7.428	7.670	7.856	8.059	8.209	8.377	8.625	8.802	2.727
(5):	4.229	6.234	6.333	6.697	7.159	7.431	7.676	7.861	8.047	8.189	8.343	8.590	8.791	2.164
(6):	4.326	6.061	6.170	6.536	7.070	7.390	7.623	7.806	7.991	8.127	8.281	8.526	8.687	1.477
(7):	4.361	6.070	6.175	6.543	7.077	7.395	7.627	7.808	7.995	8.135	8.291	8.531	8.688	1.411
(8):	4.419	6.188	6.280	6.626	7.142	7.438	7.652	7.812	7.990	8.121	8.270	8.484	8.642	0.574
(9):	4.323	6.124	6.225	6.551	7.055	7.387	7.617	7.794	7.980	8.117	8.267	8.504	8.685	1.481
(10):	4.445	6.188	6.286	6.642	7.135	7.424	7.646	7.808	7.990	8.126	8.275	8.494	8.660	0.350
(11):	4.465	6.257	6.345	6.669	7.161	7.481	7.688	7.851	8.015	8.129	8.283	8.529	8.714	1.136
ALL	4.543	6.218	6.311	6.646	7.137	7.428	7.643	7.807	7.989	8.124	8.277	8.500	8.685	0.000

Note that in Table 7, the row labeled (10) has the smallest value, which is 0.35. Thus, the representative scenarios are 186, 40, 146, 46, 124, 169, 121, 111, 180, and 112 (i.e., the row labeled (10) in Table 6).

7 The Methodology

7.1 The Statistical Tests

For each of the eight methods shown in Table 2, 1,000 scenarios are generated and then reduced to the 50 representative scenarios. This process is repeated ten times, each time using a different random number seed and the same initial yield curve. The ten repetitions of generating 1,000 scenarios for each method test the effect of the random number seed on the scenarios.

The data collected from every run consist of the following simple descriptive statistics for each of the 13 maturities: the mean, median, standard deviation, minimum, and maximum. Statistical summaries are produced for each group of 1,000 scenarios and the corresponding subsets of 50 representative scenarios as well as for the first ten years and the entire time horizon. These statistics are summarized by variable (any combination of maturity, descriptive statistic, and time period) as described in Christiansen (1994). There are 130 variables from the product of five statistics, two time periods, and 13 maturities.

The statistics for each run are summarized for each combination of method and variable as described in Christiansen (1994). These statistical summaries are compiled for all of the 10,000 scenarios and for all of the 500 representative scenarios separately. For each variable the output includes a comparison of the basic descriptive statistics of each group:

- 50 representative scenarios vs. their original 1,000 scenarios;
- 50 representative scenarios vs. all 500 representative scenarios; and
- All 500 representative scenarios vs. all 10,000 original scenarios.

A sample of the original data, which were analyzed in SAS, is included in Appendix A. Table 8 lists the statistical tests performed on the data.

The analysis of variance (ANOVA)[6] tests are used to determine stability between methods, while the univariate tests are used to test stability

[6] A good reference for ANOVA tests is Miller (1977).

Table 8
List of Statistical Tests Performed (Using SAS)

Test	Scenarios per Entry	All or Representative	Comparison	N
ANOVA	10,000	ALL (SUM)	Antithetic/Base	8
ANOVA	1000	ALL (ORG)	Antithetic/Base	20
ANOVA	500	REP (SUM)	Antithetic/Base	8
ANOVA	500	REP (SUM)	Extremes/2 Std Dev	8
ANOVA	500	REP (SUM)	Old/New Choice	8
ANOVA	50	REP	Methods	80
Univariate	1000	ALL (ORG)	Methods	-
Univariate	50	REP	Methods	-

Notes: ANOVA = Analysis of variance; ALL(SUM) = All (Summarized); N = Number of observations; ALL (ORG) = All (Original); REP = Representatives; and REP (SUM) = Representatives (Summarized)

within a method. We use the term *stability within a method* to compare similarities between values for a single variable generated with one random number seed to the same variable generated by the same method but with a different random number seed.

In SAS, the univariate procedure produces a box and whisker plot for each combination of variable and method and also places the plots created by different methods side by side to facilitate comparisons of their distributions. The range represents a worst case example of instability. The interquartile range is a more likely estimate of the instability from one trial to the next, while the mean and median give the two best point estimates of the variable under consideration.

Figure 1 illustrates a typical box and whisker plot. These plots give a quick view of the means and interquartile range of the data.[7] Half of the interquartile range can be used as an estimate of the standard deviation. The scale for the plots is determined from the stem.

7.2 Fitness Measure

Next we must develop a measure of fit to determine how close the representative group (subset of scenarios) to the overall set of scenarios. We expect the closer the fit, the more consistent the results of any process that depends on the scenarios should be.

[7]For a more detailed description see Christiansen (1998).

Figure 1
The Basic Parts of a Box and Whisker Plot

```
17         *        -> Extreme Outlier
16
15
14
13
12
11         |        -> Whiskers
10         |
 9         |
 8     +------+     -> 75th Percentile = Q3
 7     |      |
 6     *------*     -> 50th Percentile = Median
 5     |  +   |     -> Mean
 4     +------+     -> 25th Percentile = Q1
 3         |
 2         |
 1         |
 0         |
-0
-1
-2         0        -> Outlier

Stem    Boxplot
```

Not wanting the measure of fit to be skewed by the outliers, i.e., the maxima or minima in the raw data, we develop a fitness measure based on the relative error in replacing the 10,000 scenarios with the 500 representative scenarios. To avoid cancellations of one relative error with another, we consider the following measure of fit:

$$\text{FIT} = \sum_{i=1}^{2} \sum_{j=1}^{5} \sum_{k=1}^{13} w_i v_j z_k \times \left(\frac{O_{ijk} - R_{ijk}}{O_{ijk}} \right)^2 \tag{3}$$

where O_{ijk} represents the statistic based on the set of original scenarios and R_{ijk} represents the statistic based on the representative scenarios. The w_i weights reflect the relative importance of the early years of the scenario compared to all the years; the v_j refer to the relative importance of the means ($j = 1$), medians ($j = 2$), standard deviations ($j = 3$), minimum ($j = 4$), and maximum ($j = 5$); and the z_k are the relative weights of the various maturities. These weights do not need to be the same as those assigned in the choice function.

8 The Statistical Results

8.1 Differences Between Methods

Three levels of statistical results are provided. The first level of results provided is the values of the mean, median, standard deviation, minimum, and maximum for each of the 13 maturities. These values serve as the inputs to the second level. The second level is the collection of ANOVA tests that show significance for a particular variable at a particular level. Because large numbers of these tests are performed, a third level of statistical comparisons must be made before any significance can be attributed to the comparisons made. Comparisons are made using the significance of each group of results from the analyses of variance performed on the 130 independent variables.

A simple estimate of the significance of results is obtained by the normal approximation to the binomial random variable; see Devore (1982, p. 201). For the group of 130 tests and a five percent significance level, the normal approximation has mean 6.5 (130×0.05) and standard deviation 2.485 ($\sqrt{130 \times 0.05 \times (1 - 0.05)}$). Thus for a group of results to be significant at the 5 percent level, at least 11 of the individual results need to be significant at that level. For groups of results that have multiple levels of significance a more accurate estimation of the probability that the results are due to chance is obtained from a multinomial distribution, where the results are separated into two to four non-overlapping categories depending on the level of significance.

The results of the individual analyses of variance comparing antithetic normal random variables to the base case (no antithetic normal) show significant differences at the 5 percent level for the following six variables:

- The seven year median (all years);
- One year mean (years 0–10);
- Two year mean (years 0–10);
- 15 year standard deviation (years 0–10);
- 20 year standard deviation (years 0–10); and
- 30 year standard deviation (years 0–10).

No variable is significantly different at the 1 percent level. This result is consistent with an expectation of 6.5 significant results at the 5 percent level due to chance when running 130 tests. Therefore, this group of results is not significant.

For the original data (groups of 1,000) there are no statistically significant differences due to the use of antithetic random deviates. Moreover, for some of the variables there are no differences. This is significantly fewer than would be expected by chance. The results of the ANOVA tests show that only a single variable is significant at the 5 percent level. This is consistent with the lack of significance for the antithetic versus base random variable test on the original scenarios.

The analysis of variance shows highly significant differences, as expected, between the groups of 500 representative scenarios depending upon the choice function. All of the maturities display significant differences for the first year means. Table 9 summarizes the differences. Each decimal entry is the probability that the differences are due to chance. Only those that are significant are entered; a dash indicates lack of significance at the 5 percent level.

The significance of the results of Table 9 is determined from the probabilities associated with a multinomial distribution, i.e.,

$$\begin{aligned} p &= \binom{130}{5\,6\,1\,9\,109}(0.04)^5(0.009)^6(0.0009)^1(0.0001)^9(0.95)^{109} \\ &= 2.611 \times 10^{-28}. \end{aligned}$$

There are 109 results that are not significant $(130-(5+6+1+9))$ with an associated probability of 0.95.

As expected, the analysis of variance shows highly significant differences between the groups of 500 representative scenarios depending upon whether the candidate list is determined using extremes or the two standard deviation method. Table 10 summarizes these results. From the multinomial distribution we obtain

$$\begin{aligned} p &= \binom{130}{12\,13\,3\,5\,97}(0.04)^{12}(0.009)^{13}(0.0009)^3(0.0001)^5(0.95)^{97} \\ &= 6.72 \times 10^{-29}. \end{aligned}$$

The methods that choose the candidate list by extremes tend to find more of the extremes in their representative scenarios than do the methods that emphasize the two standard deviations. Because the minimums permitted by the MCP process are relatively close to the current rate level, there is little observed difference in minimum values.

ANOVA tests, which compare the methods, look at the total effects of the combinations of the various modifications, including the synergistic effects as well as those due to individual modifications.

Table 9
Significance Levels for Differences Between Old and New Choice Functions

Maturity	Mean	Median	Std. Dev.	Min
Three Month F	0.0046^a	0.0280	0.0032^a	-
Six Month F	0.0033^a	0.0421	0.0046^a	-
One Year F	0.0004^b	-	-	-
Two Years F	0.0001^c	-	0.0258	-
Three Years F	0.0001^c	-	0.0480	-
Four Years F	0.0001^c	-	-	-
Five Years F	0.0001^c	-	-	-
Seven Years F	0.0001^c	-	-	-
Ten Years F	0.0001^c	-	-	0.0267
15 Years F	0.0001^c	-	-	-
20 Years F	0.0001^c	-	-	-
30 Years F	0.0001^c	-	-	-
Shape A	0.0022^a		0.0011^a	

Notes: The variable names ending with "A" refer to the all years (total time horizon), while those with "F" refer to the first year's time horizon. Entries significant at the 5 percent but not at the 1 percent level are listed without superscripts. ^{a} denotes an entry that is significant at the 1 percent but not at the 0.10 percent level. ^{b} denotes an entry that is significant at the 0.10 percent but not at the 0.01 percent level. ^{c} denotes an entry that is significant at the 0.01 percent level. "-" indicates lack of significance at the 5 percent level.

Table 11 summarizes the significant results for the analyses of variance performed on the representative data comparing methods. From the multinomial distribution we have

$$
\begin{aligned}
p &= \binom{130}{14\;13\;9\;47\;47}(0.04)^{12}(0.009)^{13}(0.0009)^{9}(0.0001)^{47}(0.95)^{47} \\
&= 1.16\times 10^{-188}.
\end{aligned}
$$

Table 10

Differences Between the Extremes Method And the Two Standard Deviation Method

Variable	Median	Std. Dev.	Max	Min
Three Month A	0.0008^b	0.0001^c	0.0439	-
Six Month A	0.0001^c	0.0001^c	0.0413	-
One Year A	0.0001^c	0.0001^c	0.0260	-
Two Year A	0.0005^b	0.0004^b	0.0189	-
Three Year A	0.0036^a	0.0012^a	0.0365	-
Four Year A	0.0038^a	0.0026^a	0.0288	-
Five Year A	0.0036^a	0.0046^a	0.0345	-
Seven Year A	0.0069^a	0.0069^a	-	-
10 Year A	0.0065^a	0.0089^a	-	-
15 Year A	0.0032^a	0.0128	-	-
15 Year F	-	-	-	0.0148
20 Year A	0.0031^a	0.0332	-	-
20 Year F	-	-	-	0.0034
30 Year A	0.0214	-	-	-
30 Year F	-	-	-	0.0357

Notes: The variable names ending with "A" refer to the all years (total time horizon), while those with "F" refer to the first year's time horizon. Entries significant at the 5 percent but not at the 1 percent level are listed without superscripts. ^{a} denotes an entry that is significant at the 1 percent but not at the 0.10 percent level. ^{b} denotes an entry that is significant at the 0.10 percent but not at the 0.01 percent level. ^{c} denotes an entry that is significant at the 0.01 percent level. "-" indicates lack of significance at the 5 percent level.

Table 11
Significance of Methods

Variable	Mean	Median	Std. Dev.	Max	Min
Three Month A	-	0.0001^{c}	0.0001^{c}	0.0001^{c}	-
Three Month F	-	-	0.0005^{b}	0.0139	-
Six Month A	-	0.0001^{c}	0.0001^{c}	0.0001^{c}	-
Six Month F	-	-	0.0002^{b}	0.0115	-
One Year A	-	0.0001^{c}	0.0001^{c}	0.0001^{c}	-
One Year F	0.0234	-	0.0001^{c}	0.0059^{a}	-
Two Year A	-	0.0001^{c}	0.0001^{c}	0.0001^{c}	-
Two Year F	0.0160	-	0.0001^{c}	0.0205	-
Three Year A	-	0.0001^{c}	0.0001^{c}	0.0001^{c}	-
Three Year F	0.0092^{a}	-	0.0001^{c}	0.0334	0.0034^{a}
Four Year A	-	0.0001^{c}	0.0001^{c}	0.0001^{c}	-
Four Year F	0.0084^{a}	-	0.0001^{c}	0.0380	0.0001^{c}
Five Year A	-	0.0001^{c}	0.0001^{c}	0.0001^{c}	-
Five Year F	0.0087^{a}	-	0.0001^{c}	0.0390	0.0002^{b}
Seven Year A	-	0.0001^{c}	0.0001^{c}	0.0001^{c}	
Seven Year F	0.0066^{a}	-	0.0001^{c}	0.0490	0.0003^{b}
Ten Year A	-	0.0001^{c}	0.0001^{c}	0.0001^{c}	-
Ten Year F	0.0062^{a}	-	0.0001^{c}	0.0486	0.0008^{b}
15 Year A	-	0.0001^{c}	0.0001^{c}	0.0001^{c}	0.0056^{a}
15 Year F	0.0037^{a}	-	0.0001^{c}	-	0.0015^{a}
20 Year A	0.0310	0.0001^{c}	0.0003^{b}	0.0001^{c}	-
20 Year F	0.0095^{a}	-	0.0001^{c}	-	0.0015^{a}
30 Year A	0.0224	0.0001^{c}	0.0010^{b}	0.0001^{c}	0.0163
30 Year F	0.0214	-	0.0001^{c}	-	0.0041^{a}
Shape A	0.0001^{c}	-	0.0001^{c}	-	-
Shape F	0.0014^{b}	-	0.0006^{b}	-	-

Notes: The variable names ending with "A" refer to the all years (total time horizon), while those with "F" refer to the first year's time horizon. Entries significant at the 5 percent but not at the 1 percent level are listed without superscripts. The superscript ^{a} denotes an entry that is significant at the 1 percent but not at the 0.10 percent level; ^{b} denotes an entry that is significant at the 0.10 percent but not at the 0.01 percent level; and ^{c} denotes an entry that is significant at the 0.01 percent level. A "-" indicates lack of significance at the 5 percent level.

To confirm that the results observed in the multinomial distribution are due to synergistic effects of the three modifications (antithetic normal, candidate list, and choice function) interacting with each other, we used a likelihood ratio test. The synergistic effects hypothesis was tested against the alternative hypothesis of independence. The likelihood ratio is 70.8707, which has a probability of 1.4864×10^{-14} based on a χ^2 distribution with 4 degrees of freedom. Thus the synergistic effects are highly significant.

8.2 Differences Within a Method

We use three ways of comparing the stability of the methods with the sets of 50 representative scenarios: (i) look at the range of each of the statistics for each of the variables, (ii) look at the standard deviations of the variables for the representative scenarios, and (iii) compare how well each of the methods fits the original 1000. The range of results is selected because it is a worst case example of instability. If the first run produces the maximum for any variable and the second run produces the minimum, the instability is the difference between the two runs, i.e., the range. Box and whisker plots for the means illustrate the stability of variables.

The SAS univariate procedure is run on each of the variables that provides numeric data on the range and the box and whisker plots. All of the plots in the text are for the seven year maturity. Although all of the tests are performed on all years and all data, the examples from the univariate procedure are limited to the seven year rate. For each method these plots compare the statistical summary of each set of 50 representatives from one seed to the others. The more consistent the results from one seed to another seed, the smaller the range will be. Figure 2 illustrates the differences in the ranges of the single most important variable for the crediting rate determination: the mean of the seven year spot rates over the years zero through ten, inclusive.

Figure 2 shows that the mean of the means of the seven year rates varies by method. Methods 2, 3, and 4 display a higher mean than the other methods. But our primary concern is the range of results. We would like the range to be small and tight. Methods 5, 7, and 8 display the largest range, followed by method 6. This agrees with our intuition because the original methods do not control for the earlier years of the scenarios. Methods 2 and 4 appear to be slightly better than methods 1 and 3; although the range is the same, the interquartile distance is smaller, as evidenced by the use of a whisker instead of a box bottom.

Figure 2
Differences in the Ranges of the Mean Spot Rates
Seven Year Maturity: First Ten Years, Inclusive

```
Variable=MEAN

  8.5 +
      |
      |                                      |
      |    *---*  *-+-*  +-+-+  *-+-*  +---+      |    +---+  +---+
  8.0 +    +-+-+    |    *---*    |    *-+-*  *-+-*  *-+-*  *-+-*
      |                                +---+    0    +---+  +---+
      |                                                |      |
      |
  7.5 +
       -----+-----+-----+-----+-----+-----+-----+-----+--
METHOD      1     2     3     4       5    6     7     8
```

The mean of the seven year rate over the first years is not the only variable with which we are concerned, although it may be the most important. Figure 3 displays the mean over all years of the seven year rate.

Based on the original requirements for this variable (one assigned the highest weight), all of the methods perform well. They all had the same mean of means and interquartile distances, although methods 7 and 8 have outliers, and method 1 has whiskers. This figure illustrates the strength of the method, but does not help with the current problem. The mean over all years is the only statistic that is controlled in the original choice function, while the new choice function adds consideration of the means in the first years of the scenarios.

Because none of the other statistics is controlled by the choice function, we confine our discussion to the early years because these years are more important than the overall scenario. Figure 4 compares the medians of the seven year rate by method.

The medians appear to have a smaller range than the means, with method 2 appearing to be the best and method 6 to be the worst. Method 5 has the largest interquartile range of any of the methods. Again, the pattern is that those methods that attempt to control the means for the first years have a narrower range in general than those that only control the means over the scenarios as a whole.

The comparison of methods for standard deviations of the seven year rate for the first years of the scenarios is found in Figure 5.

Figure 3
Seven Year Maturity: Mean Over All Years

```
Variable=MEAN

  8.5 +
      |
      |       |                                                      0         0
      |     *---*    *-+-*    *-+-*    *-+-*    *-+-*    *-+-*    *-+-*    *-+-*
  8.0 +       |
      |
      |
  7.5 +
       -----+-----+-----+-----+-----+-----+-----+-----+--
METHOD      1     2     3     4     5     6     7     8
```

Ideally, the distribution of the standard deviations would be compact. While these are less compact than the distribution of the means and medians, the scale is twice as sensitive. Method 5, has the lowest whiskers. The mean and median coincide for methods 1, 2, 4, 5, and 7. The interquartile distance on method 2 is smaller than that of method 7; however method 2 has much longer whiskers and the overall distribution is larger. Methods 3, 4, and 8 tend to be higher than the others, with whiskers that go up rather than down. This may be desirable for cash flow testing, but is probably not as desirable in setting a crediting rate strategy.

The relative stability of the methods (or ranking) varies by which statistic we are considering (mean, median, standard deviation, minimum, or maximum) as well as by maturity rate. Data for the first ten years inclusive are summarized by maturity in the following tables which were extracted from the detailed information from the SAS univariate procedure used to create the comparative box and whisker plots. Appendix B contains a sample of the output from the SAS univariate procedure. Each of the tables is arranged, by maturity, in order of the decreasing range of results for each method. The rankings do not indicate the relative variation in the ranges. In Table 12, which considers the means, the largest range is for method 7 (41.9 basis points, Appendix B), and the smallest range (2.4 basis points, Appendix B) is for method 2. Generally for all of the maturities methods 5 and 7 produce larger ranges than do methods 2 or 4. The rankings for the means are remarkably consistent from variable to variable.

Figure 4
Comparison of Methods for Medians
Seven Year Maturity: First 10 Years, Inclusive

```
Variable=MEAN

          |
          |
   8.0 +
          |      *-*-*  *-*-*  *-+-*  *-*-*  *-+-*  *-+-*  *-+-*  *-*-*
          |        *             *      *    +---+    |      0      *
          |                                           0
          |
          |
   7.5 +
          -----+-----+-----+-----+-----+-----+-----+-----+--
METHOD         1     2     3     4     5     6     7     8
```

For the medians (see Table 13), the results are not as consistent from variable to variable, although method 2 consistently has a smaller range than does method 7. For many of the variables the differences are small. The ranges for the seven year maturity are from 27.9 basis points for method 6, to 15.9 for method 7, to 10.1 for method 2.

For the standard deviation (see Table 14), there is a reversal in rankings for the seven year maturity. Method 7 has a smaller range of standard deviations (38.2 basis points) than does method 2 (54.9 basis points), while method 3 has the largest range of standard deviations (72.2 basis points).

These statistics have in one way or another reflected all of the data, while the minimum and maximum reflect the impact of a single number. The following discussion is limited to the shorter time horizon. For the minima, the range of results by maturity is given in decreasing order in Table 15. The minima, especially on the short end of the curve, are influenced by the absolute minimum permitted rate in a scenario. This creates the larger candidate lists from which to choose, and (with the current level of rates) virtually guarantees that all of the runs will have the absolute minimum as the minimum (leaving a range of 0). Once we move away from the shortest maturities this situation no longer exists, and there are definite differences by method. The range for the minima for the seven year maturity for method 2 is 1.009, while for method 7 it is 0.514.

Figure 5
Comparison of Methods for Standard Deviations
Seven Year Maturity: First 10 Years, Inclusive

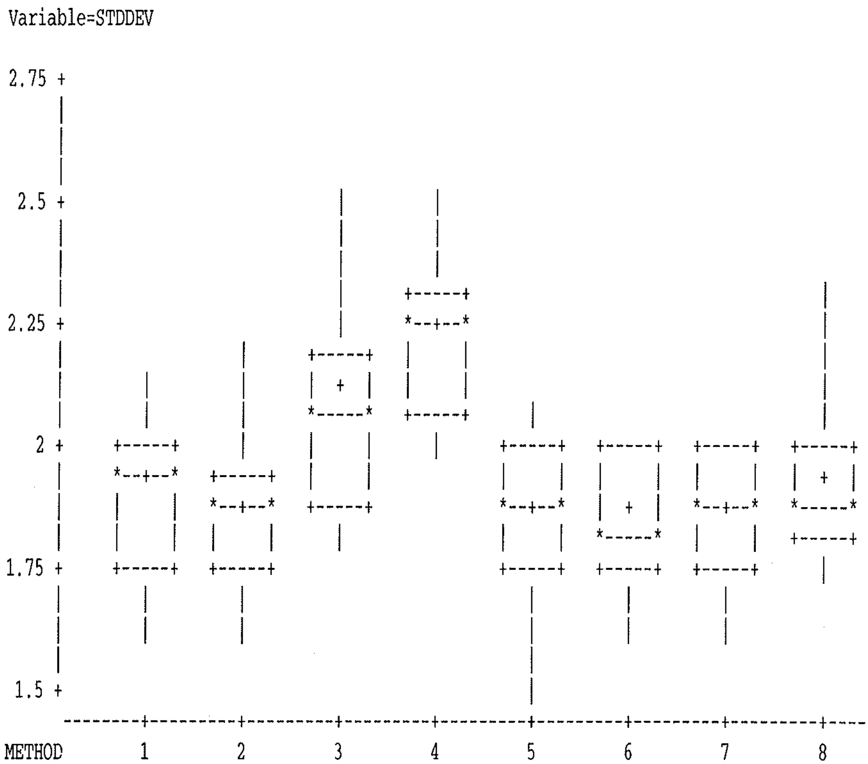


Table 16 only gives the relative rankings and does not indicate the size of the ranges of the maxima. The range of the maxima for the seven year maturity for method 2 is 8.886 percent, while for method 7 it is 2.675 percent. While these ranges seem huge, their importance is tempered by the fact that we are comparing the worst outliers of two runs of the same method starting with the same curve.

A different way of looking at the stability issue is to examine the standard deviations for each of the variables. The smaller the standard deviations produced, the more stable the results. In this case we look at the mean (and dispersion) of the standard deviations to see where we can expect the lowest variability or, conversely, the most predictability. This is not the same as examining the variables labeled standard deviations, but is an examination of the standard deviations of each of the original statistics for the 130 variables. Actual values can be obtained from the complete SAS univariate procedure data.

Table 17 illustrates the results of the following weightings in the fit measure. The entries in the columns headed *mean, median, std dev, min,* and *max* are the results of the intermediate calculations which consider the maturity and first years/all years weightings. The column headed *chooser* is the sum of the previous columns.

In this example all of the maturities are weighted equally. For these combinations of weights, method 2 appears to be best and method 4 the worst. Other examples are given in Appendix C. Preliminary testing of a switch from method 7 to method 2 indicates that such a switch solves the 40 basis point problem.

Table 12

Ranking of Methods by Range of Means

Maturity	Largest			To		Smallest		
Three Month	5	7	8	3	4	6	1	2
Six Month	5	7	8	3	6	4	2	1
One Year	7	5	8	3	6	4	1	2
Two Year	5	7	8	3	6	1	2	4
Three Year	5	7	8	6	3	1	4	2
Four Year	7	5	8	6	3	1	4	2
Five Year	7	5	8	6	3	1	4	2
Seven Year	7	5	8	6	3	1	4	2
Ten Year	7	8	5	6	3	1	4	2
15 Year	8	5	7	6	3	1	4	2
20 Year	8	5	7	6	3	1	4	2
30 Year	8	7	5	6	3	1	4	2

Table 13
Ranking of Methods by Range of Medians

Maturity	Largest			To				Smallest
Three Month	7	1	5	4	2	8	3	6
Six Month	7	2^a	4^a	3^a	5^b	1^b	8	6
One Year	7	3	1	5	8	2	4	6
Two Year	5	7	1	4^a	2^a	6^a	3^a	8^a
Three Year	5	6	7	1	3	8	2	4
Four Year	5	6	7	4	3	8^a	1^a	2^a
Five Year	5	8	7	6	4	2	3	1
Seven Year	6	3	5	8	7	1	4	2
Ten Year	6	5	3	8	7	1	4	2
15 Year	5	7^a	8^a	4	6	3	1	2
20 Year	8	7	5	6	1	4	3	2
30 Year	8	5	6	7	1	2	4	3

Notes: [a] means the ranges are the same to three decimal places; [b] means the ranges are the same to three decimal places; but not the same as [a] for the same maturity.

Table 14
Ranking of Methods by Range of Standard Deviations

Maturity	Largest			To				Smallest
Three Month	1	5	3	4	8	2^a	6^a	7
Six Month	1	5	3	4	8	6	7	2
One Year	1	3	5	8	6	4	7	2
Two Year	3	5	1	8	6	4	2	7
Three Year	3	5	8	4	1	6	2	7
Four Year	3	5	8	6	4	1	2	7
Five Year	3	5	8	4^a	2^a	1	6	7
Seven Year	3	5	2	8	4	1	6	7
Ten Year	3	5	2	4	8	1	6	7
15 Year	3	4	8	1	2	5	6	7
20 Year	3	4	2	5	8	1	6	7
30 Year	3	4	2	8	1	5	6	7

Notes: [a] denotes same ranges, to three decimal places.

Table 15
Ranking of Methods by Range of Minima

Maturity	Largest			To				Smallest
Three Month	-	-	-	-	-	-	-	-
Six Month	-	-	-	-	-	-	-	-
One Year	-	-	-	-	-	-	-	-
Two Year	5	6	1	-	-	-	-	-
Three Year	6	5	2	1	4	8	7[a]	3[a]
Four Year	5	6	2	1	8	4	3	7
Five Year	2	5	6	1	3	4	8	7
Seven Year	5	2	6	1	3	4	8	7
Ten Year	5	2	6	1	8	3	7	4
15 Year	2	8	6	5	7	1	3	4
20 Year	1	7	3	2	6	4	8	5
30 Year	1	2	3	7	6	4	8	6

Notes: [a] denotes same ranges, to three decimal places; and – denotes no difference.

Table 16
Ranking of Methods by Range of Maxima

Maturity	Largest			To				Smallest
Three Month	8	5	4	1	6	3	7	2
Six Month	8	5	4	1	6	3	7	2
One Year	4	1	8	6	5	3	7	2
Two Year	4	1	5	8	2	7	3	6
Three Year	4	2	5	8	1	6	3	7
Four Year	5	2	4	8	6	1	3	7
Five Year	5	4	2	6	8	1	3	7
Seven Year	5	2	4	6	8	1	3	7
Ten Year	2	5	4	6	8	1	3	7
15 Year	5	2	4	8	1	3	6	7
20 Year	2	8	4	6	1	5	3	7
30 Year	2	8	1	3	6	4	5	7

Table 17
Sample of Fit Measures

Method	Mean	Median	Std	Min	Max	Chooser
1	0.0007	0.0976	2.9160	0.0462	0.1038	3.1643
2	0.0004	0.0860	1.5746	0.1391	0.0889	1.8890
3	0.0013	0.0443	6.6476	0.0000	1.7590	8.4523
4	0.0012	0.0331	11.8840	0.0000	0.1760	12.0943
5	0.0183	0.2325	2.4490	0.3619	0.4949	3.5566
6	0.0255	0.1130	2.1062	0.1371	1.1959	3.5778
7	0.0120	0.0462	0.9522	0.0000	3.2169	4.2274
8	0.0198	0.0729	1.5756	0.0518	2.7586	4.4787

9 Conclusions and Caveats

We have shown that there are no significant differences due to the use of antithetic normal random variables when one considers either the original scenarios or the representative scenarios. The use of antithetic random variables without any other change was not statistically significant; but their inclusion led to very highly significant synergistic effects. We also have established that both the method of determining the candidate list and the choice function are highly significant. The differences between methods are extremely significant, showing that there are synergistic effects as well as those effects due to the original three modifications we perform.

Differences within a method were not determined due to the lack of a separate objective standard for comparison (and also a small sample size). The main impetus for the study is the difference within a method from one trial to the next. The differences within a method do become manageable for the seven year mean (the single most important variable) when the method is changed from the original method (7) to the method using all three of the modifications (2).

We introduce a fit measure, a generalized least squares percentage measure of the error introduced by using the representative scenarios instead of the entire set of scenarios. This measure also supports the replacement of method 7 with method 2.

Because it is necessary from a practical point of view to run only a representative sample of interest rate scenarios rather than 10,000 or even 1,000, it is necessary to choose the method that will be the most acceptable on an overall basis. Therefore, consideration must be given to the relative strengths and weaknesses of each method in light of the application planned for the scenarios. These measures can only provide limited guidance and a suggestion about which method to use to determine the representative 50 scenarios. Stability may be enhanced by moving from scenarios based on method 5 or 7 to scenarios based on method 2 which incorporates all of these proposed variance-reducing techniques.

There are still many unanswered questions. Does the shape of the original curve impact the stability of the method? If the stability is defined by the range of results and 1000 scenarios are not sufficiently stable, how many scenarios are necessary before they are reduced to the representative scenarios?

An application of the univariate procedure to the groups of 10,000 scenarios is not a measure of variation from seed to seed, as each of the summarized groups of 10,000 scenarios has ten different seeds. If the scenarios were determined from several distinct seeds, would that improve the stability? These scenarios were determined from a model in which the key rate was the ten year rate. Would scenarios designed for cash flow testing and based on a 30 year key rate have different results?

References

Bratley, P., Fox, B.L., and Schrage, L.E. *A Guide to Simulation.* New York, N.Y.: Springer-Verlag, 1983.

Christiansen, S.L.M. "The Markov Chain Interest Rate Generator Revisited." *Journal of Actuarial Practice,* 2, no. 1 (1994): 101–124.

Christiansen, S.L.M. "Representative Interest Rate Scenarios." *North American Actuarial Journal* 2, no. 3 (1998): 29–59.

Devore, J.L. *Probability and Statistics for Engineering and the Sciences.* Monterrey, Calif.: Brooks Cole, 1982.

Kalos, M.H. and Whitlock, P.A. *Monte Carlo Methods Volume 1: Basics.* New York, NY.: Wiley and Sons, 1986.

Miller R.B. and Wichern, D.W. *Intermediate Business Statistics.* New York, NY.: Holt Rinehardt and Winston, 1977.

Ostaszewski, K. *An Investigation into Possible Applications of Fuzzy Set Methods in Actuarial Science.* Ithasca, Ill.: Society of Actuaries, 1993.

Robbins, E.L., Cox, S.H. and Phillips, R.D. "Applications of Risk Theory to Interpretation of Stochastic Cash Flow Testing Results." *North American Actuarial Journal* 1, no. 2 (1997): 85-104.

Tilley, J. "An Actuarial Layman's Guide to Building Interest Rate Generators." *Transactions of the Society of Actuaries* 44 (1987): 509-538.

Appendix A—Sample Data

Table A1
Summarized Data Groups of 10,000

Variable	Mean	Median	Std. Dev.	Min	Max	Method
Seven Year f	8.058	7.924	1.764	3.500	19.921	1
Seven Year f	8.076	7.924	1.796	3.500	22.919	2
Seven Year f	8.073	7.924	1.774	3.500	23.018	3
Seven Year f	8.073	7.924	1.772	3.500	21.967	4
Seven Year f	8.074	7.924	1.776	3.500	21.814	5
Seven Year f	8.074	7.924	1.778	3.500	20.832	6
Seven Year f	8.072	7.924	1.763	3.500	21.310	7
Seven Year f	8.078	7.924	1.805	3.500	25.000	8

Table A2
Summarized Data 500 Representatives

Variable	Mean	Median	Std. Dev.	Min	Max	Method
Seven Year f	8.059	7.924	1.917	3.500	19.499	1
Seven Year f	8.073	7.924	1.897	3.500	22.919	2
Seven Year f	8.075	7.924	2.109	3.500	19.121	3
Seven Year f	8.074	7.924	2.225	3.500	21.967	4
Seven Year f	7.986	7.796	1.861	3.606	21.428	5
Seven Year f	7.960	7.912	1.848	3.528	19.851	6
Seven Year f	7.993	7.893	1.867	3.500	16.636	7
Seven Year f	7.983	7.924	1.950	3.526	19.582	8

Table A3
Data for Each Individual Group of Representatives
Methods 2 and 7

Variable	Mean	Median	Std. Dev.	Min	Max	Method
Seven Year f	8.081	7.823	2.195	4.066	22.919	2
Seven Year f	8.074	7.904	2.067	3.805	16.329	2
Seven Year f	8.064	7.924	1.726	3.500	18.492	2
Seven Year f	8.071	7.924	1.646	4.136	14.391	2
Seven Year f	8.076	7.876	1.964	4.127	17.565	2
Seven Year f	8.082	7.924	1.899	4.470	17.457	2
Seven Year f	8.084	7.924	1.876	3.606	16.379	2
Seven Year f	8.060	7.924	1.907	3.669	22.076	2
Seven Year f	8.070	7.924	1.886	4.024	18.396	2
Seven Year f	8.071	7.924	1.764	4.509	14.033	2
Seven Year f	8.137	7.924	1.779	3.735	14.153	7
Seven Year f	7.764	7.858	1.729	3.568	14.910	7
Seven Year f	8.183	7.924	1.924	3.754	15.194	7
Seven Year f	7.988	7.924	1.701	3.500	13.961	7
Seven Year f	8.070	7.924	2.031	3.500	15.695	7
Seven Year f	7.949	7.883	1.987	3.719	14.873	7
Seven Year f	7.882	7.765	1.905	4.014	16.636	7
Seven Year f	8.053	7.924	1.649	3.537	14.988	7
Seven Year f	8.084	7.924	1.898	3.872	15.172	7
Seven Year f	7.824	7.893	1.992	3.500	14.744	7

Appendix B—Methods 2 and 7 Quantiles for the Seven Year Rate (Early Years) Descriptive Statistics

Table B1
Methods 2 and 7 Quantiles for the Mean and Median Of the Seven Year Rate (Early Years)

	Mean		Median	
Percentile	Method 2	Method 7	Method 2	Method 7
100% Max	8.0840	8.1830	7.924	7.924
75% Q3	8.0810	8.0840	7.924	7.924
50% Med	8.0725	8.0205	7.924	7.924
25% Q1	8.0700	7.8820	7.904	7.883
0% Min	8.0600	7.7640	7.823	7.765
Range	0.0240	0.4190	0.101	0.159
Q3-Q1	0.0110	0.2020	0.020	0.041
Mode	8.0710	7.7640	7.924	7.924

Table B2
Methods 2 and 7 Quantiles for the Minimum and Maximum Of the Seven Year Rate (Early Years)

	Minimum		Maximum	
Percentile	Method 2	Method 7	Method 2	Method 7
100% Max	4.509	4.0140	22.919	16.636
75% Q3	4.136	3.7540	18.492	15.194
50% Med	4.045	3.6435	17.511	14.949
25% Q1	3.669	3.5000	16.329	14.744
0% Min	3.500	3.5000	14.033	13.961
Range	1.009	0.5140	8.886	2.675
Q3-Q1	0.467	0.2540	2.163	0.450
Mode	3.500	3.5000	14.033	13.961

Table B3
Methods 2 and 7 Quantiles for the Standard Deviation of the Seven Year Rate (Early Years)

Percentile	Method 2	Method 7
100% Max	2.1950	2.0310
75% Q3	1.9640	1.9870
50% Med	1.8925	1.9015
25% Q1	1.7640	1.7290
0% Min	1.6460	1.6490
Range	0.5490	0.3820
Q3-Q1	0.2000	0.2580
Mode	1.6460	1.6490

Appendix C—Examples of the use of Fuzzy Choice Functions

Table C1
Example 1

Method	Mean	Median	Std Dev	Min	Max	Chooser
1	0.0002	0.0291	0.5245	0.0185	0.0415	0.6138
2	0.0001	0.0253	0.3024	0.0557	0.0356	0.4191
3	0.0003	0.0093	0.7361	0.0000	0.7678	1.5136
4	0.0003	0.0094	1.2725	0.0000	0.1054	1.3875
5	0.0039	0.0599	0.5615	0.1448	0.1980	0.9681
6	0.0054	0.0323	0.4828	0.0548	0.4784	1.0537
7	0.0026	0.0089	0.1554	0.0000	1.3206	1.4874
8	0.0041	0.0148	0.2464	0.0207	1.1036	1.3895

Notes: The weights used are as follows: 1, 1, 1, 1, 1, 1 and 2 for the mean, median, standard deviation, minimum, maximum, all years and first years, respectively.

Table C2
Example 2

Method	Mean	Median	Std Dev	Min	Max	Chooser
1	0.0003	0.0582	1.0490	0.0185	0.0415	1.1675
2	0.0002	0.0507	0.6048	0.0557	0.0356	0.7469
3	0.0006	0.0187	1.4722	0.0000	0.7678	2.2593
4	0.0005	0.0187	2.5451	0.0000	0.1054	2.6697
5	0.0079	0.1198	1.1231	0.1448	0.1980	1.5935
6	0.0107	0.0646	0.9656	0.0548	0.4784	1.5742
7	0.0052	0.0177	0.3108	0.0000	1.3206	1.6542
8	0.0082	0.0295	0.4927	0.0207	1.1036	1.6547

Notes: The weights used are as follows: 2, 2, 2, 1, 1, 1 and 2 for the mean, median, standard deviation, minimum, maximum, all years and first years, respectively.

Outlier Analysis of Annual Retail Price Inflation: A Cross-Country Study

Wai-Sum Chan*

Abstract†

Wilkie's stochastic investment model and its variants have been increasingly applied by actuaries around the world to actuarial modeling and simulation. This paper performs time series outlier analysis on retail price inflation, which is the driving force of Wilkie's composite model. The data come from four developed countries: the United Kingdom, the United States, Canada, and Australia. The fit of the model is significantly improved after the adjustment of outliers. The analysis also identifies exogenous events that have intervened in the inflation dynamics. An example is given to demonstrate the importance of outlier analysis on stochastic simulation. Finally, inflation trends for these four countries are examined. The results suggest caution in the use of the 4 percent inflation assumption of some U.K. and Australian actuaries.

Key words and phrases: *economic assumptions, inflation trend, stochastic model, time series outlier, Wilkie model*

*Wai-Sum Chan, Ph.D., F.S.A., is an associate professor of actuarial science at the University of Hong Kong. Dr. Chan has a B.B.A. (Honours) in accounting from the Chinese University of Hong Kong and a Ph.D. in applied statistics from Temple University, USA. He is a Fellow of both the Society of Actuaries and the Royal Statistical Society.

Dr. Chan's address is: Department of Statistics, The University of Hong Kong, Pokfulam Road, HONG KONG. Internet address: *chanws@hku.hk*

†Part of this research was carried out when the author was visiting the Department of Statistics and Actuarial Science, University of Waterloo, Canada. He is grateful to the University of Waterloo for providing financial and computing support. The author also is indebted to the anonymous referees for their helpful comments on an earlier version of the article.

1 Introduction

The original Wilkie (1984, 1986) model is a composite stochastic investment model that attempts to capture the interdependence of four key variables: the retail price index, share yield, share dividends, and consols yield.[1] The model has been extensively used by U.K. actuaries for various purposes, ranging from assessing the solvency of life offices (Limb et al., 1986) to modeling uncertainty in general insurance companies (Daykin and Hey, 1990). Other actuarial applications of the Wilkie model in the U.K. include Wilkie (1987), Purchase et al., (1989), Ross (1989), and Hardy (1993). Wilkie (1995) extends the original model to add five particular variables, plus a family of variables (i.e., currency exchange rates). A comprehensive review of Wilkie's model is provided by Huber (1997).

Following Wilkie's footsteps stochastic investment models have been developed for other countries. They include Metz and Ort (1993) for Switzerland; Deaves (1993) for Canada; Daykin et al., (1994) for Finland; Thomson (1996) for South Africa; Frees et al., (1997) for the U.S.; and Sherris et al., (1997) for Australia. Unfortunately, these models, as well as the original Wilkie model (see, for examples, Wilkie (1995), Kitts (1990) and Clarkson (1991)), usually produce non-normal and nonlinear residuals, which could be due to the existence of outliers in the data series.

Retaining outliers in the time series could lead to erroneous model specification and biased predictions (Chan, 1995). Chan and Wang (1998) apply the time series outlier detection technique, developed by Chen and Liu (1993), to U.K. price inflation. The results show that the residuals are significantly closer to a normal distribution. Foster (1997) also detects outliers and level shifts in U.S. real wage series. In a similar study, Balke and Fomby (1994) examine 15 U.S. macroeconomic time series and they conclude that outliers appear to be present in all series. Also, after controlling for outliers, much of the evidence of nonnormality and nonlinearity of the residuals is eliminated.

In this article we extend Chan and Wang's (1998) work to perform time series outlier analysis on price inflation, which is the most important driving force of Wilkie's composite model, for four developed countries. The price inflation series is defined by

$$I_t = \ln P_t - \ln P_{t-1} \tag{1}$$

[1] They are called the *consumer price index*, *stock return*, *stock dividends* and *long-term interest rate*, respectively, in the U.S. and Canada.

where P_t is the price index at time t. A first order autoregressive (AR(1)) model is often employed to describe the inflation dynamic:

$$I_t = \mu + \phi(I_{t-1} - \mu) + \varepsilon_t, \tag{2}$$

where for $t = 1, 2, 3, \ldots$, the ε_t (called *stochastic disturbance terms*) are independent and identically distributed (i.i.d.) normal random variables with mean 0 and variance σ^2; ϕ is the autoregressive parameter ($|\phi| < 1$); and μ is the mean rate of the inflation process.

This model is widely accepted by actuaries for pension simulations and other actuarial applications (see, for example, Knox (1993) and Wilkie (1995)). The model can be interpreted as follows: each year the force of inflation (I_t) is equal to its mean rate (μ), plus a proportion (ϕ) of last year's excess inflation ($I_{t-1} - \mu$), plus a random disturbance (ε_t) which has zero average and variance σ^2.

The main objectives of this paper are:

- To show actuaries the importance of outlier analysis in building a stochastic model;
- To identify global exogenous events that might have significant impact on the inflation dynamic of different countries;
- To study the inflation trend for each country under examination.

2 Outlier Analysis

2.1 Time Series Outlier Models

Time series observations are often influenced by interruptive events such as strikes, outbreaks of wars, sudden political or economic crises, or even unnoticed errors of typing and recording. The consequences of these interruptive events create aberrant observations, which are inconsistent with the rest of the series. Such observations are usually referred to as *outliers*. Most outliers are not simply spurious observations (e.g., recording or typing errors). They may contain important information about the external interruptive events affecting the series. In general, outliers in time series can be viewed as the result of non-repetitive interventions. Thus, a contaminated inflation series I_t^* consists of an outlier-free inflation series I_t plus an exogenous intervention effect $\Delta_t(T, \omega)$, i.e.,

$$I_t^* = I_t + \Delta_t(T, \omega) \tag{3}$$

where I_t follows the model equation (2), T is the location of the outlier, and ω is the magnitude of the outlier.

Four commonly used types of outliers (see Tsay, 1988) and two newly proposed types of outliers (De Jong and Penzer, 1998) are considered in this paper: additive outlier (AO), innovational outlier (IO), level shift (LS), temporary change (TC), switch outlier (SO), and linear increase outlier (LIO).

- An additive outlier affects only the level of the given observation.
- An innovational outlier affects all observations beyond the given time through the memory of the underlying outlier-free process.
- A level shift is an event that affects a time series at a particular time point whose effect becomes permanent.
- A temporary change is an event having an initial impact and whose effect decreases exponentially according to a fixed dampening parameter, say, δ. In practice the value of δ often lies between 0.6 and 0.8 (Liu and Hudak, 1994, page 76). We employ $\delta = 0.7$ in this article as recommended by Chen and Liu (1993).
- A switch outlier is where there are consecutive extreme values on either side of the current level of the series. An SO would occur when the economy has a dramatic opposing change, such as from a response to unanticipated inflation or severe government intervention.
- A linear increase outlier occurs where the average level of the series ramps up to a higher level. A LIO reveals periods of regime changes within an economy such as technology improvements or again government intervention. The length of the linear increase period in a LIO is denoted by q, and we assume $q = 4$ in this study.

The form of $\Delta_t(T, \omega)$ for each type of outlier is given as:

$$
\begin{aligned}
AO: \quad & \Delta_t(T,\omega) &=& \quad \omega D_t^{(T)} \\
IO: \quad & \Delta_t(T,\omega) &=& \quad \frac{\omega}{1-\phi B} D_t^{(T)} \\
LS: \quad & \Delta_t(T,\omega) &=& \quad \frac{\omega}{1-B} D_t^{(T)} \\
TC: \quad & \Delta_t(T,\omega) &=& \quad \frac{\omega}{1-\delta B} D_t^{(T)} \\
SO: \quad & \Delta_t(T,\omega) &=& \quad \omega \times \left(D_t^{(T)} - D_t^{(T+1)}\right) \\
LIO: \quad & \Delta_t(T,\omega) &=& \quad \omega \times \left[\sum_{k=0}^{q-1} \left(\frac{k+1}{q+1}\right) D_t^{(T+k)} + \frac{1}{1-B} D_t^{(T+q)}\right]
\end{aligned}
$$

where B is the backward shift operator such that $B^s D_t^{(T)} = D_{t-s}^{(T)}$, and

$$
D_t^{(T)} = \begin{cases} 1 & \text{if } t = T \\ 0 & \text{if } t \neq T \end{cases} \tag{4}
$$

is the indicator variable representing the presence or absence of an outlier at time T. Graphical examples of the $\Delta_t(T,\omega)$ function for various types of outlier are given in Figure 1.

More generally, a time series may contain m outliers of different types, and we have the following general time series outlier model:

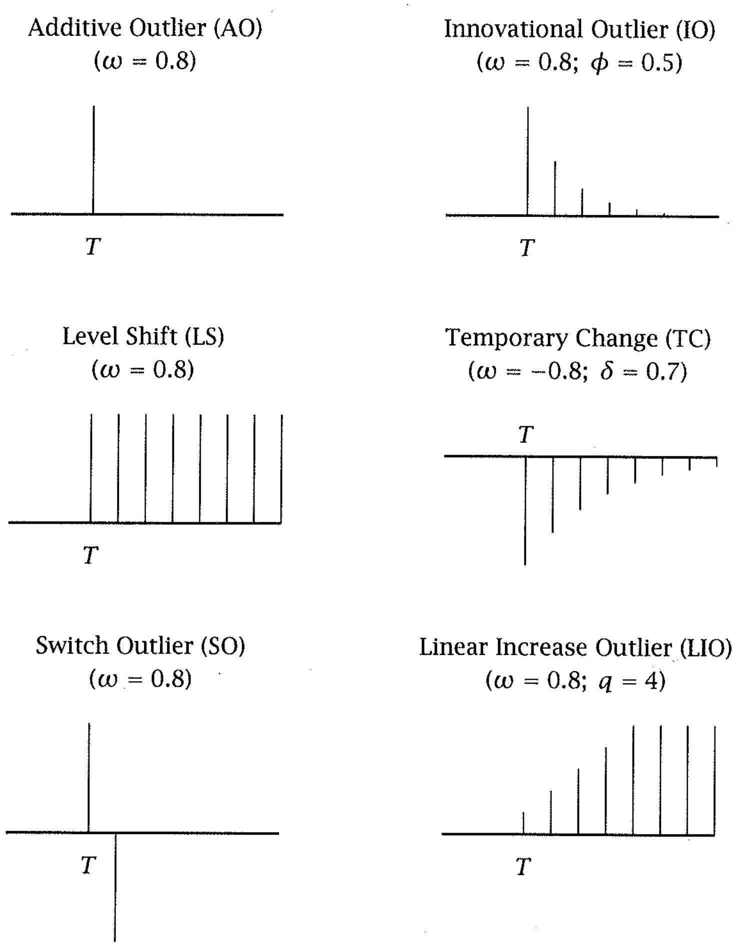
$$
I_t^* = I_t + \sum_{j=1}^{m} \Delta_t(T_j, \omega_j). \tag{5}
$$

In this article we assume that the underlying outlier-free process for I_t is AR(1). For other time series outlier models with underlying outlier-free process following a general autoregressive moving average model, see Tsay (1988); or for a general state-space model, see De Jong and Penzer (1998).

2.2 Outlier Detection and Adjustment

The search for the location and type of an outlier in a contaminated time series is known as an *outlier detection problem* in time series literature. It was first studied by Fox (1972), who employed the likelihood ratio test. Chang, Tiao, and Chen (1988) extend Fox's idea and propose an iterative procedure to detect multiple outliers. Chen and Liu (1993) further develop a simultaneous estimation and outlier detection procedure. Their approach consists of three-stage iterative cycle based on detection, estimation, and adjustment.

Figure 1
Effects of Time Series Outliers



Chen and Liu's (1993) method is used in this paper. For the detection stage, the fitted residuals $\hat{\varepsilon}_t^* = (I_t^* - \hat{I}_t)$ from model equation (2) are first obtained. The outlier effects in model equation (3) will be transmitted from the contaminated time series to the fitted residuals. Therefore, time series regressions can be written as

$$\hat{\varepsilon}_t^* = \omega \; d(s,t) + \varepsilon_t \quad for \; s \in S, \tag{6}$$

where $S = \{\text{AO}, IO, LS, TC, SO, LIO\}$ is the set of residual types;

$$\begin{aligned}
d(s,t) &= \begin{cases} 0 & \text{for all } s \text{ and } t < T \text{, and} \\ 1 & \text{for all } s \text{ except, } s = \text{L}IO\text{, and } t = T; \end{cases} \\
d(\text{L}IO, T) &= \frac{1}{q+1}; \\
d(\text{AO}, t) &= \begin{cases} -\hat{\phi} & \text{for } j = 1 \\ 0 & \text{for } j \geq 2; \end{cases} \\
d(\text{IO}, t) &= 0 \quad \text{for all } j; \\
d(\text{L}S, t) &= 1 - \hat{\phi} \quad \text{for all } j; \\
d(\text{TC}, t) &= \delta^{j-1}(\delta - \hat{\phi}); \\
d(SO, t) &= \begin{cases} -(1+\hat{\phi}) & \text{for } j = 1, \\ \hat{\phi} & \text{for } j = 2 \text{, and} \\ 0 & \text{for } j \geq 3 \end{cases} \\
d(\text{L}IO, t) &= \begin{cases} \frac{j+1-j\hat{\phi}}{q+1} & \text{for } j = 1, \ldots, q \\ 1 - \hat{\phi} & \text{for } j > q \end{cases}
\end{aligned}$$

for $t = T + j$ $(j = 1, 2, \ldots)$. They are used for detecting outliers. For given T (suspected location of the outlier) and s (suspected type of outlier), the usual regression t-statistic, $\tau(s,T)$, for the slope parameter ω in the regression model equation (5) can be computed. The final test statistic is the maximum value of this statistic searching all possible locations (T) and types (s), i.e.,

$$\mathcal{T} = \max_{1 \leq T \leq n} \max_{s \in S} \{\tau(s,T)\}. \tag{7}$$

For a given location, the test statistic follow a normal distribution approximately. An outlier is detected if $\mathcal{T}$ is greater than a critical value C. Following Liu and Hudak (1994), we employ $C = 3.5$ in this paper.

With the type and the location of an outlier, we can jointly re-estimate the model parameter and the outlier effects. After the estimation, one can adjust the outlier effects on the observations by the model equation (3). The detection-estimation-adjustment cycle is repeated for the adjusted series until no new outliers are found. Finally, the model is re-estimated for the autoregressive parameter and all outlier effects simultaneously.

The detection procedures can be easily implemented in many time series and regression computer packages. The SCA (Liu and Hudak, 1994) programming package provides outlier analysis as one of its standard features. It can automatically process four commonly used types of outliers (AO, IO, LS, and TC) in the data series. Additional macro statements can be incorporated into the system to deal with SO and LIO.

3 The Data

In this paper we apply outlier analysis to annual inflation series from 1900 to 1995 of four developed countries. They are United Kingdom, United States, Canada, and Australia.

3.1 Data Sources

Following Wilkie (1995) the U.K. Retail Prices Index for June (in each year) is taken from the following periods:

- 1900–1914: Board of Trade Wholesale Price Indices (Total Index), Table Prices 5 of Mitchell and Deane (1962);
- 1914–1947: "All Items" Cost of Living Index, Table 84 of Central Statistical Office (1991);
- 1947–1990: "All Items" Retail Prices Index, Table 1 of Central Statistical Office (1991);
- 1990–1993: "All Items" General Index of Retail Prices, Table 18.7 of Central Statistical Office (1994); and
- 1993–1995: "All Items" Retail Prices Index, Table 18.7 of Office of National Statistics (1997).

For the U.S. annual average consumer price index, two series have been combined:

- 1900–1970: Consumer Price Index, U.S. Department of Commerce (1973).
- 1970–1995: Consumer Price Index Number, International Monetary Fund (1998).

Note that Consumer Price Index Number has variable name AL64 in the International Financial Statistics Database.

For Canada, two annual average consumer price index series have been connected:

- 1900–1914: Price Indexes of Selected Retail Services, Statistics Canada (1965).
- 1914–1995: "All Items" Consumer Price Index, Statistics Canada, CANSIM Database.

Note that the CANSIM Database can be accessed through the internet at: *http://www.statcan.ca/english/CANSIM/index.html.* The price data are stored in matrix 9957 and it costs Cdn$3.00 per series.

For Australia, annual average retail price index numbers from 1900 to 1995 are recorded in Australian Bureau of Statistics (1997, p. 660).

3.2 Data Description

We use 1923 as the common base year for all index series. They are shown in Figure 2 using a vertical logarithmic scale. The graphs have similar shape. Their corresponding inflation series, as defined in equation (1), are plotted in Figure 3.

Descriptive statistics for the inflation series are summarized in Table 1. The U.K. and Australia, on the average over the past 96 years, have one percent inflation rate per year higher than the U.S. and Canada. The standard deviation in U.K. inflation is higher than that of the other countries investigated. The distributions of inflation for Canada and Australia are negatively skewed. The inflation distributions appear to have thick right tails, while Australian inflation is closer to a normal distribution.

Skewness (κ_1) and excess kurtosis (κ_2) of a random variable X are:

$$\kappa_1 = \frac{\mu_3}{\mu_2^{3/2}} \quad \text{and} \quad \kappa_2 = \frac{\mu_4}{\mu_2^2} - 3$$

where $\mu_k = E[(X - E[X])^k]$.

Table 1
Summary Statistics for Inflation Series

	U.K.	U.S.	Canada	Australia
Mean	0.0409	0.0307	0.0313	0.0408
S.D.	0.0678	0.0491	0.0481	0.0532
Skewness	0.2814	0.1057	-0.1822	-0.1577
Kurtosis	1.7796	1.4538	1.7748	0.9324
r_1	0.5200	0.6200	0.5800	0.5500
r_2	0.3200	0.2500	0.3100	0.2900
		Correlation		
U.K.	1.00			
U.S.	0.67	1.00		
Canada	0.76	0.91	1.00	
Australia	0.66	0.64	0.74	1.00

Note: S.D. = Standard deviation; r_1 = First-lag autocorrelation; r_2 = Second-lag autocorrelation.

The evidence of highly significant first-lag autocorrelation coefficients plus the fact that all r_2 are approximately equal to the square of their corresponding r_1 in Table 1 support the use of the AR(1) model for the inflation dynamic for each country. The correlations among inflation series are high. Due to geographic, political, and economic reasons, it is not a surprise to observe that the correlation between U.S. inflation and Canadian inflation is 0.91, the highest among all other combinations.

4 Empirical Results

4.1 Model Fitting

A first-order autoregressive model is fitted to each data series without considering the possibility of outlier effects. Based on the outlier analysis described in Section 2, AR(1) models are also fitted to the outlier-adjusted series. Table 2 presents fitting results under both situations.

Figure 2
Retail (Consumer) Price Indexes, 1990–1995

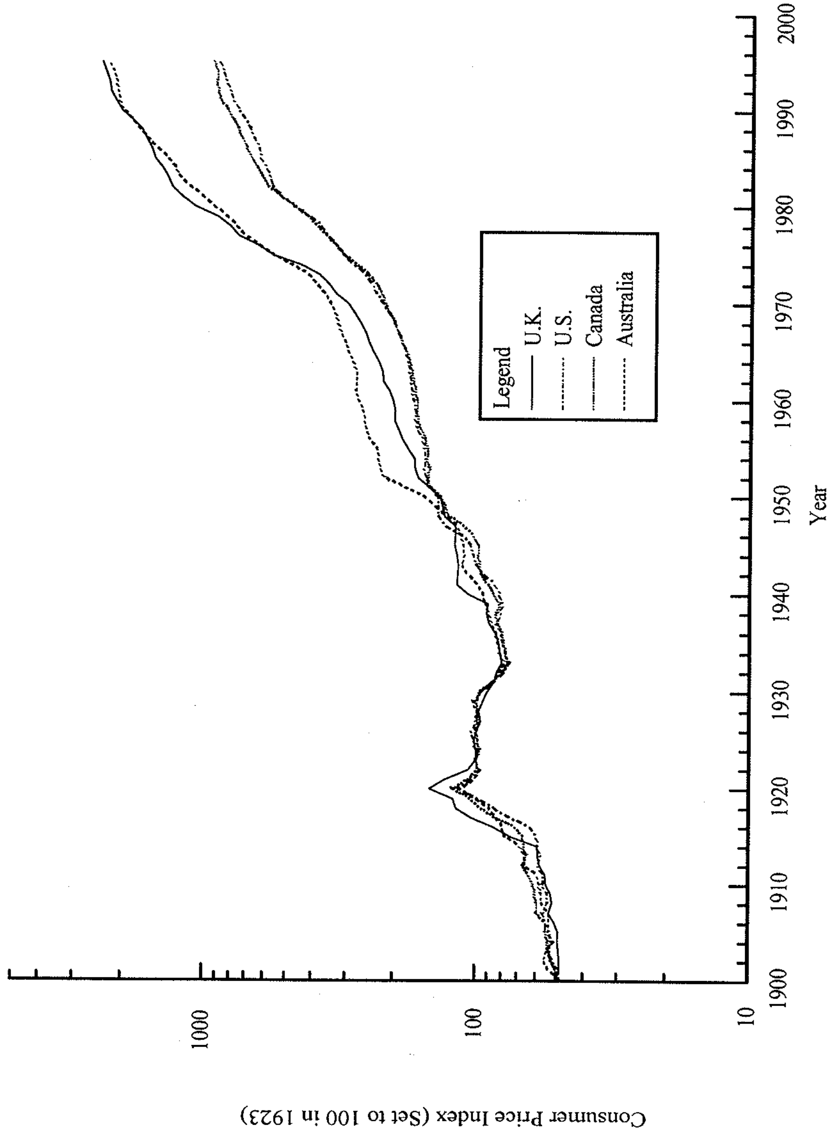


Figure 3
Price Inflation Series, 1990-1995

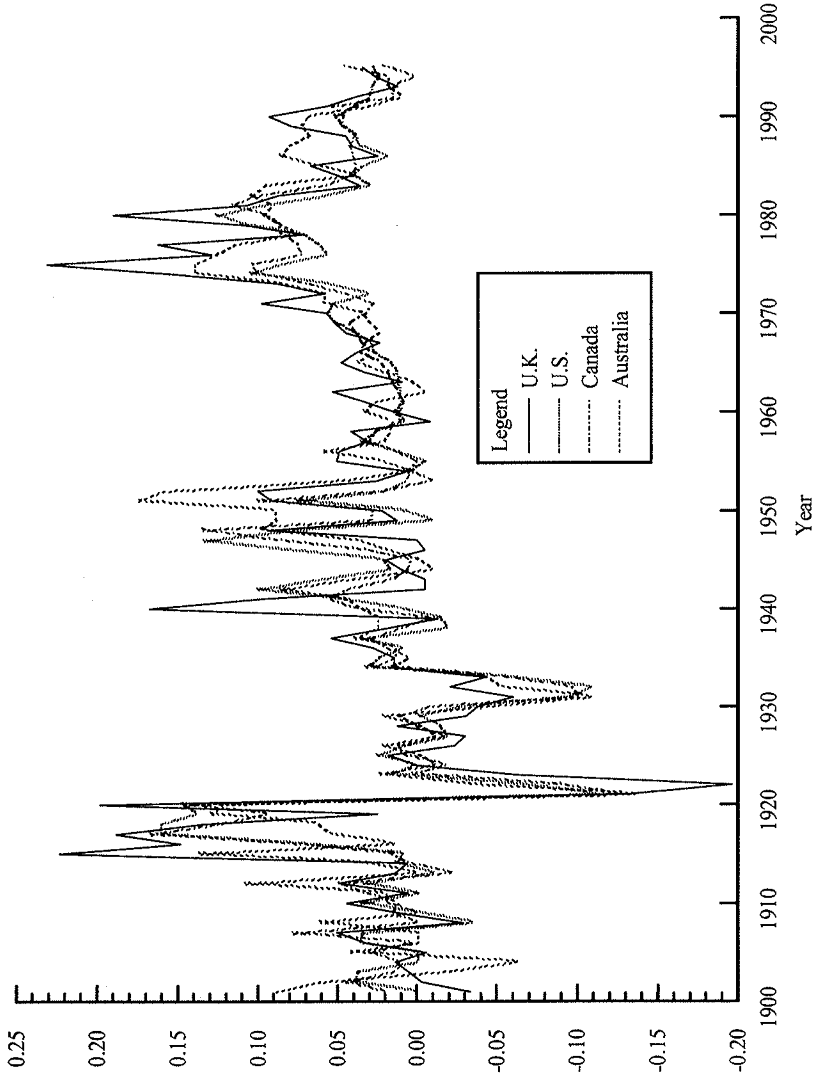


Table 2
Model Fitting Results for the Inflation AR(1) Process of Equation (1) For Various Countries (Before and After the Outlier Adjustments)

	United Kingdom		United States		Canada		Australia	
	Before	After	Before	After	Before	After	Before	After
$\hat{\mu}$	0.042	0.029	0.032	0.040	0.031	0.040	0.040	0.046
$\hat{\phi}$	0.519	0.534	0.617	0.752	0.575	0.547	0.550	0.692
$\hat{\sigma}$	0.058	0.035	0.039	0.031	0.039	0.027	0.044	0.035
AIC[1]	-529.3	-612.3	-603.0	-644.9	-601.2	-665.1	-577.7	-618.3
AIC(2)	-520.8	-607.4	-596.6	-641.7	-590.8	-664.9	-564.0	-609.4
JB[2]	103.6	3.1	354.8	2.3	463.5	3.6	141.9	2.7
NLT[3]	0.012	0.447	0.325	0.987	0.604	0.789	0.114	0.310
Q[4]	15.7	10.6	10.6	12.3	8.1	4.7	11.7	4.2

[1]AIC = $n \ln \hat{\sigma}^2 + 2M$, where n is the number of effective observations, and M is the number of parameters in the model. Under this criterion, we should choose the model with *smaller* AIC. AIC(2) is the AIC value for an alternative AR(2) model for the series.

[2]JB is the Jarque and Bera's (1981) test statistic for normality of the residuals. Under the null (normality) hypothesis, the critical value of the test is 5.99 at the 5 percent level.

[3]NLT is the p-value of Tsay's (1989) F test for linearity of the residuals. We will reject the null (linearity) hypothesis if the p-value is less than the significance level, say, 5 percent.

[4]Q is the Ljung and Box's (1978) Portmanteau statistic (with 10 lags) for testing serial correlation of the residuals. Under the null (independence) hypothesis, the critical value of the test is 15.507 at the 5 percent level.

The residual standard deviation ($\hat{\sigma}$) is significantly reduced for each country after adjusting for outliers. Because we have introduced additional parameters into the outlier model, it is not appropriate to focus only on the improvement of the fit. Akaike (1974) proposes an information criterion to compare alternative models fitted to a dataset with different number of parameters. The criterion has been called AIC (Akaike Information Criterion) in the literature and is defined as

$$AIC(M) = n \ln(\hat{\sigma}^2) + 2M \tag{8}$$

where n is the number of effective observations and M is the number of parameters in the model. The criterion considers both the model fitting ($\hat{\sigma}^2$) and the model parsimony (M). Under this criterion, one should choose the model with the smallest AIC. The results in Table 2 indicate that the outlier model is preferred to the original model in every country. They all give a smaller value of AIC. To guard against model misspecification, the AIC value for an alternative AR(2) process is computed for each country. The results justify our choice of the AR(1) model for the inflation series.

The problem of nonnormality and nonlinearity of residuals from the original AR(1) model for U.K. inflation has caused some concerns for many authors (see, Kitts, 1990; Geoghegan et al., 1992; Huber, 1997; and Chan and Wang, 1998). In this paper we examine the normality and linearity of the residuals using the Jarque and Bera (1981) test and the Tsay (1989) test, respectively. The results in Table 2 show that normality of the residuals has been remarkably improved after controlling for outliers. Furthermore, degree of nonlinearity in the residuals could also be alleviated by the outlier model in each country. Finally, Portmanteau Q statistics (Ljung and Box, 1978) are computed for testing serial correlation of the residuals. The results do not indicate any inadequacy of the fitted models.

4.2 Detected Outliers

Table 3 displays the outliers found for each country. It describes the type, size, and t-ratio of the outlier as well as the year in which it occurred. In addition, we also try to link the year of each outlier to an economic event that occurred in or near that year.

An examination of Table 3 reveals that British inflation is more vulnerable to external shocks compared to other countries. Five outliers are identified. On the other hand, the U.S. inflation is more robust to economic disturbances. Only one outlier is detected in 1921.

Table 3
Detected Outliers in Chronological Order

		United Kingdom			United States		
Year	Event	Type	Size	*t*-ratio	Type	Size	*t*-ratio
1915	World War I	TC	0.208	6.30			
1917	World War I						
1920	Post-WWI	SO	0.160	4.77			
1921	Post-WWI				TC	-0.233	-7.54
1922	Post-WWI	AO	-0.188	7.22			
1931	Recession (Canada)						
1940	World War II	IO	0.163	4.83			
1975	Oil Crisis Shock	TC	0.139	4.25			

		Canada			Australia		
Year	Event	Type	Size	*t*-ratio	Type	Size	*t*-ratio
1915	World War I				AO	0.112	3.87
1917	World War I	TC	0.109	4.09			
1920	Post-WWI						
1921	Post-WWI	TC	-0.229	-8.57	IO	-0.235	-6.64
1922	Post-WWI						
1931	Recession (Canada)	TC	-0.119	-4.46			
1940	World War II						
1975	Oil Crisis Shock						

In addition to the global events, the Canadian inflation dynamic was interrupted by the internal severe economic depression in early 1930s (*cf.* Dominion Bureau of Statistics, 1938, p. 813). Historically, the inflation dynamic in Australia was disturbed twice (1915 and 1921), possibly due to the effects of World War I.

5 An Example

The importance of outlier analysis in stochastic (actuarial) simulation will now be demonstrated.

Consider an insurance company that is interested in selling index-linked policies.[2] Before offering such policies, the company's actuaries would need to consider the characteristics of index-linked assets to match the resulting index-linked liabilities.[3]

In this example we consider the U.K. government indexed bonds, which are commonly called *index-linked gilts.*[4] For simplicity, we assume that the bond makes annual coupon payments that are based on the inflation-adjusted face value of the bond over time. The initial face amount is 1,000, coupon interest rate $c = 5$ percent per annum, and time to maturity $N = 20$ years. The adjustment for inflation is made using the annual U.K. Retail Price Index (RPI) with a one year lag. At maturity, the redemption value also is adjusted for the realized inflation between the initial indexation year and one year prior to the maturity. We further assume that the bond is currently selling at par (i.e., $P = 1,000$).

Let y be the yield rate for the bond; y can be obtained by solving the following bond pricing equation:

$$P = \sum_{t=1}^{N} \frac{cF_t}{(1+y)^t} + \frac{F_N}{(1+y)^N} \tag{9}$$

where F_t is the inflation-adjusted face value at time t. The current value of F_t is the last year's face value (F_{t-1}) adjusted by the lagged inflation, i.e.,

[2] Wilkie (1981) presents arguments in favor of inflation-indexed life insurance contracts.

[3] Recently, several countries have started issuing inflation-indexed government bonds, that is, securities with yields that rise and fall with inflation. Such bonds provide tools for matching index-linked liabilities. For more details of inflation-indexed bonds, see Huh (1996).

[4] The U.S. version of government indexed bonds is called *Treasury Inflation-Protection Securities* (TIPS), see Roll (1996) and Madsen (1998) for details.

$$F_t = F_{t-1}e^{I_{t-1}} \tag{10}$$

with $F_0 = 1,000$.

If the force of inflation is assumed to be static at the 4 percent level (i.e., $I_t = I = 0.04$), the yield rate (y) can be solved analytically as

$$\begin{aligned} y &= (1+c)e^I - 1 \\ &= (1+0.05)e^{0.04} - 1 = 9.285\%. \end{aligned}$$

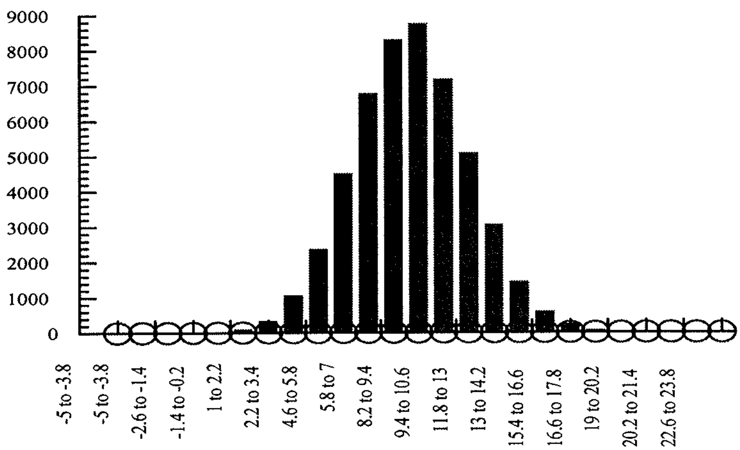
If I_t follows an AR(1) process, the distribution of y can be studied through simulation. Twenty years of inflation rates are generated using the fitted AR(1) process in Table 2 (without outlier adjustments). Given the inflation rates, the value of y can be solved from equation (8). The experiment is repeated 50,000 times. The simulation study is also carried out using the fitted AR(1) model after adjusting the outliers in Table 2. The empirical distributions of y, under both situations, are plotted in Figure 4. After controlling the outliers, actuaries are able to obtain a more precise distribution of y.

Finally, we should emphasize that the sole objective of this example is to demonstrate the reduction of the volatility of the yield of an inflation-protected bond under traditional pricing methods. We do not mean that one should price these bonds using non-risk neutral expected pricing techniques. Discussions on pricing considerations of these index-linked gilts are, however, beyond the scope of this paper.

6 Inflation Trends

Following Chan and Wang (1998), we study the inflation trend for each country. Periods from starting year (SY) through 1995 are considered. The starting year is rolled from 1900 to 1971. It creates 72 periods. The last period 1971–1995 has 25 observations which is the minimum for computing reasonable AR(1) estimates. In each period we calculate the mean rate of the inflation process ($\hat{\mu}$) after controlling the outliers. We repeat the computation for each country. The results are plotted in Figure 5. There is an upward trend in the long-term mean of the U.K. inflation process: it climbs from 4 percent to 8 percent. There is also an upward trend in the long-term mean of the Australia inflation process.

Figure 4
Simulated Distribution of y (Yield Rate) Hypothetical U.K. Inflation-Indexed Bond



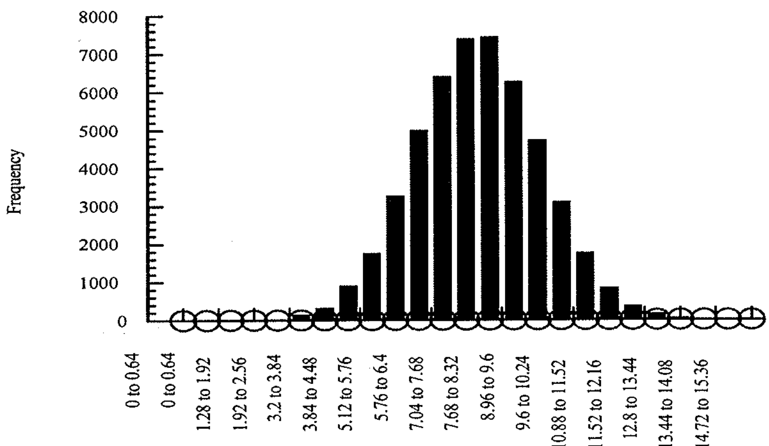
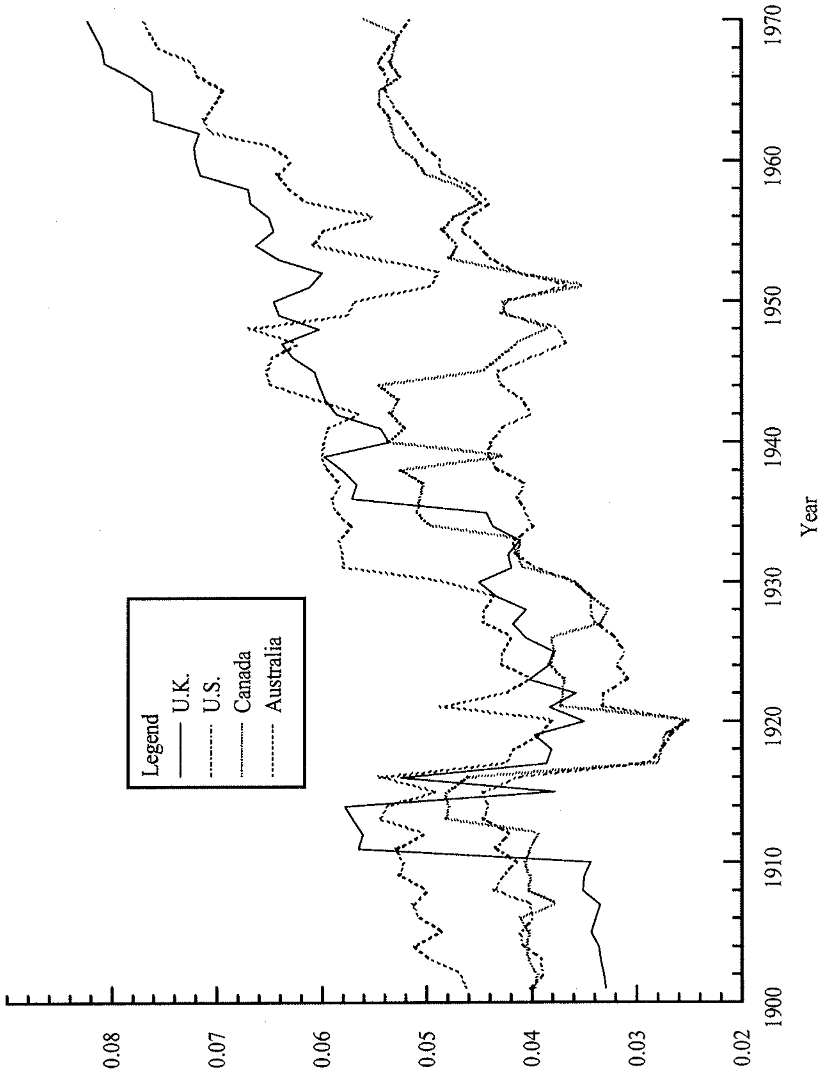


Figure 5
Inflation Trends



The results urge caution the use of 4 percent inflation assumption made by some U.K. and Australian actuaries (see, for example, Cooper, 1997, p. 18; and Knox, 1993, p. 54). On the other hand, inflation trends for the U.S. and Canada seem to fluctuate around the 4 percent level.

7 Conclusion

This research highlights the importance of outlier analysis for actuaries wishing to construct and use stochastic investment models. We perform time series outlier analysis on price inflation, which is a driving force of most existing actuarial investment models. Several exogenous events that have intervened in the inflation dynamics are identified.

The U.K. outlier-adjusted inflation model is applied to examine the distribution of the yield rate of an index-linked gilt. Finally, inflation trends are studied. The results question the use of 4 percent inflation assumption by some U.K. and Australian actuaries.

In addition to price inflation, stochastic modeling of interest rates, investment yields and other component variables in the Wilkie model are also important in a life office (see, for examples, Bragg, 1984; Greeley and Leff, 1984; Panjer and Bellhouse, 1980; and Smart, 1977). Actuaries are reminded to include outlier analysis as an initial step for modeling these variables.

References

Akaike, H. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* AC-19, (1974): 716-723.

Australian Bureau of Statistics (ABS). *Year Book Australia.* Canberra, Australia: ABS, 1997.

Balke, N.S. and Fomby, T.B. "Large Shocks, Small Shocks, and Economic Fluctuations: Outliers in Macroeconomic Time Series." *Journal of Applied Econometrics* 9 (1994): 181-200.

Bragg, J.M. "Inflation Rates and Investment Yields: The Interrelationships." *Transactions of the 22nd International Congress of Actuaries* 3 (1984): 69-80.

Central Statistical Office (CSO). *Retail Prices, 1914-1990.* London, England: Her Majesty's Stationary Office (HMSO), 1991.

Central Statistical Office (CSO). *Annual Abstract of Statistics.* London, England: Her Majesty's Stationary Office (HMSO), 1994.

Chan, W.S. "Outliers and Financial Time Series Modelling: A Cautionary Note." *Mathematics and Computers in Simulation*, 39 (1995): 425-430.

Chan, W.S. and Wang, S. "The Wilkie Model for Retail Price Inflation Revisited." *British Actuarial Journal* 4 (1998): 637-652.

Chang, I., Tiao, G.C. and Chen, C. "Estimation of Time Series Parameters in the Presence of Outliers." *Technometrics* 30 (1988): 193-204.

Chen, C. and Liu, L.M. "Joint Estimation of Model Parameters and Outlier Effects in Time Series." *Journal of the American Statistical Association* 88 (1993): 284-297.

Clarkson, R.S. "A Non-linear Stochastic Model for Inflation." *Transactions of the 2nd AFIR International Colloquium* 3 (1991): 233-253.

Cooper, D.R. "Providing Pensions for U.K. Employees with Varied Working Lives." *Journal of Actuarial Practice* 5 (1997): 5-41.

Daykin, C.D. and Hey, G.B. "Managing Uncertainty in a General Insurance Company." *Journal of the Institute of Actuaries* 117 (1990): 173-277.

Daykin, C.D., Pentikäinen, T. and Pesonen, M. *Practical Risk Theory for Actuaries.* London, England: Chapman and Hall, 1994.

De Jong, P. and Penzer, J. "Diagnosing Shocks in Time Series." *Journal of the American Statistical Association* 93 (1998): 796-806.

Deaves, R. *Modelling and Predicting Canadian Inflation and Interest Rates.* Canada: Canadian Institute of Actuaries, 1993.

Department of Commerce. *Long Term Economic Growth.* Washington, DC: U.S. Department of Commerce, 1973.

Dominion Bureau of Statistics. *The Canadian Year Book.* Ottawa, Canada: Department of Trade and Commerce, 1938.

Foster, R.S. Discussion of "Forecasting Social Security Actuarial Assumptions." *North American Actuarial Journal* 1 (1997): 79-81.

Fox, A.J. "Outliers in Time Series." *Journal of the Royal Statistical Society* B34 (1970): 350-363.

Frees, E., Kung, Y.C., Rosenberg, M., Young, V. and Lai, S.W. "Forecasting Social Security Actuarial Assumptions." *North American Actuarial Journal* 1 (1997): 49-77.

Geoghegan, T.J., Clarkson, R.S., Feldman, K.S., Green, S.J., Kitts, A., Lavecky, J.P., Ross, F.J.M., Smith, W.J. and Toutounchi, A. "Report on the Wilkie Stochastic Investment Model." *Journal of the Institute of Actuaries* 119 (1992): 173-228.

Greeley, C. and Leff, H.B. "Reserves and Solvency in a Fluctuating Interest Rate Environment." *Transactions of the 22nd International Congress of Actuaries* 3 (1984): 245-254.

Hardy, M.R. "Stochastic Simulation in Life Office Solvency Assessment." *Journal of the Institute of Actuaries* 120 (1993): 131-151.

Huber, P.P. "A Review of Wilkie's Stochastic Asset Model." *British Actuarial Journal* 3 (1997): 181-210.

Huh, C. "Inflation-Indexed Bonds." *Risks and Rewards* (The Newsletter of the Investment Section of the Society of Actuaries) no. 25 (1996): 23-24.

International Monetary Fund. *International Financial Statistics Database* (CD-ROM). Washington, D.C.: IMF, 1998.

Jarque, C.M. and Bera, A.K. "An Efficient Large Sample Test for Normality of Observations and Regression Residuals." *Working Papers in Economics and Econometrics* No. 47 Australian National University, 1981.

Kitts, A. "Comments on a Model of Retail Price Inflation." *Journal of the Institute of Actuaries* 117 (1990): 407-412.

Knox, D.M. "A Critique of Defined Contribution Plans Using a Simulation Approach." *Journal of Actuarial Practice* 1 (1993): 49-68.

Limb, A.P., Hardie, A.C., Loades, D.H., Lumsden, I.C., Mason, D.C., Pollock, G., Robertson, E.S., Scott, W.F. and Wilkie, A.D. "The Solvency of Life Assurance Companies." *Transaction of Faculty of Actuaries* 39 (1986): 251-340.

Liu, L.M. and Hudak, G.B. *Forecasting and Time Series Analysis Using the SCA Statistical System.* Chicago, Ill.: Scientific Computing Associates, 1994.

Ljung, G.M. and Box, G.E.P. "On a Measure of Lack of Fit in Time Series Models." *Biometrika* 65 (1978): 297-303.

Madsen, C.K. "Inflation-Protected Securities—A Further Look." *Risks and Rewards* (The Newsletter of the Investment Section of the Society of Actuaries) no. 30 (1998): 12-15.

Metz, M. and Ort., M. "Stochastic Models for the Swiss Consumer's Price Index and the Cost of the Adjustment of Pensions to Inflation for a Pension Fund." *Transactions of the 3rd AFIR International Colloquium* 2 (1993): 789-806.

Mitchell, B.R. and Deane, P. *Abstract of British Historical Statistics.* Cambridge, London: Cambridge University Press, 1962.

Office of National Statistics (ONS). *Annual Abstract of Statistics.* London, England: Her Majesty's Stationary Office (HMSO), 1997.

Panjer, H.H. and Bellhouse, D.R. "Stochastic Modeling of Interest Rates with Applications to Life Contingencies." *The Journal of Risk and Insurance* 47 (1980): 91-110.

Purchase, D.E., Fine, A.E.M., Headdon, C.P., Hewitson, T.W., Johnson, C.M., Lumsden, I.C., Maple, M.H., O'Keeffe, P.J.L., Pook, P.J. and Robinson, D.G. "Reflections on Resilience: Some Considerations of Mismatching Tests, with Particular Reference to Non-linked Long-term Insurance Business." *Journal of the Institute of Actuaries* 116 (1989): 347-452.

Ross, M.D. "Modelling a With-Profits Life Office." *Journal of the Institute of Actuaries* 116 (1989): 691-715.

Roll, R. "U.S. Treasury Inflation-Indexed Bonds: The Design of a New Security." *Journal of Fixed Income* 5 (1996): 9-28.

Sherris, M., Tedesco, L. and Zehnwirth, B. "Stochastic Investment Models: Unit Roots, Cointegration, State Space and GARCH Models for Australia Data." *ARCH* No. 1 (1997): 95-144.

Smart, I.C. "Pricing and Profitability in a Life Office." *Journal of the Institute of Actuaries* 104 (1977): 215-158.

Statistics Canada. *Historical Statistics of Canada.* Ottawa, Canada: Statistics Canada, 1965.

Thomson, R.J. "Stochastic Investment Modelling: The Case of South Africa." *British Actuarial Journal* 2 (1996): 765-801.

Tsay, R.S. "Outliers, Level Shifts, and Variance Changes in Time Series." *Journal of Forecasting* 7 (1988): 1-20.

Tsay, R.S. "Testing and Modeling Threshold Autoregressive Processes." *Journal of the American Statistical Association* 84 (1989): 231-240.

Wilkie, A.D. "Indexing Long-Term Financial Contracts." *Journal of the Institute of Actuaries* 108 (1981): 299-360.

Wilkie, A.D. "Steps Towards a Comprehensive Stochastic Investment Model." *Occasional Actuarial Research Discussion Paper* London, England: Institute of Actuaries, 1984.

Wilkie, A.D. "A Stochastic Investment Model for Actuarial Use." *Transactions of Faculty of Actuaries* 39 (1986): 341-403.

Wilkie, A.D. "Stochastic Investment Models—Theory and Applications." *Insurance: Mathematics and Economics* 6 (1987): 65-83.

Wilkie, A.D. "More on a Stochastic Asset Model for Actuarial Use." *British Actuarial Journal* 1 (1995): 777-964.

Journal of Actuarial Practice Vol. 6, 1998

An Analysis of Australian Pensioner Mortality by Pre-Retirement Income

David Knox* and Andrew Tomlin†

Abstract‡

The existence of a relationship between an individual's socioeconomic status and his or her mortality is often accepted, but it is difficult to measure this relationship objectively. This study analyses the relationship between an individual's final salary immediately prior to retirement and mortality rates during retirement. The data used are taken from a large Australian public sector pension plan. A strong inverse relationship is found, which decreases with age. Some of the implications of these results for individual annuity markets and public pension policy are discussed.

Key words and phrases: *salary, public pension plan, annuity, pension policy*

*David Knox, Ph.D., is the Foundation Professor of Actuarial Studies and Director of the Centre for Actuarial Studies at the University of Melbourne. He has previously taught at Macquarie University, Australia and the University of Waterloo, Canada. Professor Knox has acted as a consultant to a range of organizations, including life offices, merchant banks, and the Australian Federal Treasury. Prior to joining academia, he worked for a major life office. He has written and spoken extensively on topics related to the development of Australian superannuation during the last decade or so. His particular interest is the design features of Australia's retirement income system. Professor Knox is currently Vice President of The Institute of Actuaries of Australia and was named "Actuary of the Year" in 1996. He is also a member of the Board of the Australian Prudential Regulation Authority.

Professor Knox's address is: Centre for Actuarial Studies, University of Melbourne, Parkville, Victoria 3052, AUSTRALIA. Internet address: *d.knox@ecomfac.unimelb.edu.au*

†Andrew Tomlin is a research assistant in the Centre of Actuarial Studies.

Mr. Tomlin's address is: Department of Economics, University of Melbourne, Parkville, Victoria 3052, AUSTRALIA.

‡We wish to thank Mr. Peter Agnew of the Australian Government Actuary's Office within the Insurance and Superannuation Commission for his work in preparing the data for this project. We also acknowledge financial support received from the Australian Research Council-Small Grant No. S11947411. We also appreciate the comments made by the referees.

1 Introduction

Most retirement income plans around the world (whether they be public plans, occupational plans, or personal pension plans) provide retirees with a lifetime pension payable, totally or in part, from the assets accumulated during the individual's working career.[1] In many instances these pensions are paid until the death of the individual retiree or his/her spouse.

The life expectancies of all pensioners within a particular plan are not the same and are affected by a number of factors. There is considerable evidence (see, for example, Carney and Hanks, 1994 and World Bank, 1994) that mortality rates are linked to the socioeconomic status of the individual, which may be measured by occupation, wealth, lifetime income, education, or a combination of these factors. The specific relationships between longevity and socioeconomic factors are difficult to measure due to the lack of longitudinal data.

The objective of this study is to further our understanding in this area by testing the relationship between the mortality rates of retirees from a particular pension plan and their pre-retirement salary levels. All members of this plan must accept a lifetime pension, so there is no opportunity for selection or opting out by those with higher or lower life expectancies. This is an important difference from many previous studies[2] that have considered annuity markets (as a whole) where some individual choice (and hence selection) exists in terms of either the level or type of the annuity chosen or whether to participate in the market.

The possible link between socioeconomic status and mortality is also important in both the design and equity considerations of public pension plans, as these plans have important redistributive functions within a society. Some writers (e.g., World Bank, 1994 and Atkinson, Creedy, and Knox, 1996) have raised the issue of intragenerational equity involved in public pension plans due to the links between an individual's lifetime income and mortality. As a result, it has been suggested (for example, in the World Bank Report) that the provision of a lifetime pension in public retirement income plans may introduce inequity, as those with longer lifetimes receive pensions which have a present value that exceed the accumulated value of their contributions.

There are different issues within public and private sector plans. In public plans all individuals may be eligible for a pension that is paid from contributions and taxes received from a variety of sources. These

[1] Retirement income plans are called retirement income schemes in some countries.

[2] For example, the Continuous Mortality Investigation Reports from the United Kingdom.

public arrangements reflect social and political decisions made within a particular society at a particular time. Even in these plans, however, it is feasible that the presence of differential mortality may lead to financial redistribution through the public pension arrangements that works in a manner contrary to other principles normally adopted by the society. For instance, significant regressivity may occur, whereas progressivity is often a feature of tax and social security systems.

In contrast to public plans, privately purchased individual annuities are priced to take into account a number of significant factors that are known to affect mortality and that can be practically used (for example, age and gender). There are normally other significant factors (for example, ethnicity) that cannot be used for a number of reasons. In many cases, this approach means that annuity providers will assume, for very good reasons, that selection will occur, thereby making the annuity market less attractive to some potential investors.

In considering both public and personal arrangements, a fundamental question exists as to the extent and significance of any relationship between a socioeconomic factor (say, income) and mortality. If higher income individuals have a lower expected mortality rate (and hence higher life expectancy), then this factor needs to be considered in the design of public pension plans, the funding of occupational pensions and, if practical, the determination of annuity prices. Of course, the recognition and measurement of such a relationship does not mean that there is an easy practical solution in any of these situations to the dilemma that differential mortality may present.

This paper assesses the relationship between pre-retirement salary and post-retirement mortality for individuals from a large public sector pension plan in Australia, where all members must receive a lifetime pension related to their period of service and final salary.

2 The Data and Methodology

2.1 The Data

An investigation into the links between lifetime income and mortality after retirement requires income data and mortality records for many individuals over many years. Ideally, a longitudinal study would be conducted comparing lifetime income and post-retirement mortality. Such records are virtually impossible to obtain, however, as they would require detailed recording over a 50, 60, or 70 year period. Furthermore, within the Australian context, most retirement benefits are taken in a lump sum form, so even the mortality experience of pensioners is difficult to obtain. With these limitations in mind, this study has concentrated on the data available from one of the largest pension plans in the country, namely the Commonwealth Superannuation Scheme, a pension plan for Australian federal public servants.

This plan provides a pension related to the member's final salary and completed years of service for all members who retired prior to June 30, 1994. No commutation (substitution) of the employer-funded pension is permitted, thus enabling each retiree's final salary to be calculated from the current indexed pension and the dates that each individual joined and retired from the plan. It may be preferable for the investigation to use lifetime income. This measure is unavailable. As a proxy for lifetime income, the individual's final salary will be used. While this is not ideal, it represents an appropriate measure for those who have had a reasonable career in the public service. The level of final salary should provide a good indication of the level of their lifetime earnings. Manipulation of final salary to improve the individual's pension is unlikely to occur in the public sector due to industrial awards. Most of the existing pensioners are males who have been employed in the public service for at least two decades. Naturally, this gender bias among pensioners reflects past employment attitudes rather than current practice.

The raw data for this study are provided by the Australian Government Actuary and comprise records on Commonwealth Superannuation Scheme pensioners for the fiscal years ending June 30, 1991, 1992, 1993, and 1994. (Invalid pensioners are excluded from the data.) Using the data fields relating to the pensioner's date of birth, indexed pension, date joined plan, date exited plan, and date of death (if relevant), a salary at retirement can be determined for each pensioner and a mortality rate calculated for any age or income group. All salaries were converted to 1994 dollars. Further details of the methodology used for calculating salaries are given in Section 2.2.

Records of pensioners who retired before the age of 55 are ignored as these represent retrenchments where a choice of lump sum payment or pension was given such that an individual's salary can not be calculated. Pensioners with a date of retirement before June 30, 1976 also are excluded, as their pensions were calculated using a different method. Data on widowed pensioners, where a reversionary pension is paid on the death of the retired employee, are insufficient for analysis of mortality of these pensioners.

Table 1 summarizes the number of records available for each year. To calculate the pensioner mortality rates for each age and income grouping, the number of deaths and the exposure for each group are calculated (as outlined in the appendix) and summed over the four years.

Table 1
Number of Records Available for Investigation

Fiscal Year	Males	Females
July 1, 1990 to June 30, 1991	24,626	5,674
July 1, 1991 to June 30, 1992	25,886	6,106
July 1, 1992 to June 30, 1993	27,417	6,604
July 1, 1993 to June 30, 1994	28,508	7,107
Total	106,437	25,491

Pensioners with an observed annual salary at retirement of more than \$60,000 are grouped into one income band to obtain sufficient numbers of exposed lives and expected deaths for the statistical analysis. Pensioners with an annual salary of less than \$20,000 are considered likely to have been contributing to the superannuation fund while employed in a part-time capacity so that their final salary is unlikely to be a true indication of their total lifetime income. This group therefore is excluded from the analysis. (The group accounts for less than 2 percent of all males and less than 5 percent of all females.) The number of records available for pensioners age 85 and over is negligible, and these pensioners also are excluded from the study.

2.2 Calculating Indexed Pensions

The data set for each year includes a pension plan ID, the gender, date of birth and date of death (if applicable) of the pensioner, the dates she/he joined and exited the pension plan, and the pension as indexed on July 1 for that year based on the consumer price index movement from the previous March to March quarters.

To determine the total number of lives exposed to the risk of dying at each age, the age of each pensioner on July 1 each year in the four year period is recorded (age x). The fraction of the year in which the pensioner is age x and age $x + 1$ is calculated. In effect, the census method is used to calculate the exposure to risk. The total number of lives exposed to the risk of dying at each age (E_x) and the observed number of deaths at age x are summed over the four year period for each income range and each age group.

The pensioner's salary at retirement also is determined together with its equivalent value in 1994 dollars. Salaries are adjusted by consumer price index movements in the relevant years to obtain a 1994 salary figure. A wage index was considered but rejected as there was no index available for public servants. During the ten years prior to 1994 the consumer price index and the movement in average wages were both 5.4 percent per annum for the period. Salaries at retirement are calculated from the indexed pension according to the following equation:

$$\begin{aligned}\text{Indexed Pension at July 1} &= \text{Salary} \times \text{Benefit Multiple}\\ &\quad \times \text{Discount Factor} \times \text{Indexation}\end{aligned}$$

where:

- Benefit multiple is determined by a set of accrual rates corresponding to the number of years the pensioner has contributed to the pension plan;
- The discount factors are applied for early retirement; and
- Salaries are adjusted by the CPI.

3 The Results

3.1 Male Pensioners

Table 2 shows the total years of exposure (E) and the number of observed deaths (O) for each age-income group and the resulting crude central rates of mortality (M). The number of expected deaths (EXP) for each age-income grouping is also shown, based on the experienced mortality rate for the particular age group as a whole.

To check the validity of the results, the calculated mortality rates for male pensioners are compared with the corresponding mortality rates assumed for male pensioners in the Public Sector and Commonwealth Superannuation Schemes in the Australian Government Actuary's report on long-term costs using data to June 1993 (Duval, 1994).

Table 2
Mortality Experience of Male Pensioners

	50 to 59 Years				60 to 64 Years				65 to 69 Years			
Income	*E*	*O*	*M*	*EXP*	*E*	*O*	*M*	*EXP*	*E*	*O*	*M*	*EXP*
20,000–30,000	1461.2	8	0.0055	6.1	4515.1	56	0.0124	37.8	7060.8	137	0.0194	109.1
30,000–40,000	2560.1	10	0.0039	10.7	6670.3	64	0.0096	55.8	9010.3	154	0.0171	139.2
40,000–50,000	2690.4	13	0.0048	11.2	5966.8	47	0.0079	49.9	8008.7	127	0.0159	123.7
50,000–60,000	1669.0	7	0.0042	7.0	4038.1	27	0.0067	33.8	5710.9	73	0.0128	88.2
> 60,000	1454.4	3	0.0021	6.1	3803.8	15	0.0039	31.8	6197.4	65	0.0105	95.7
Total	9835.2	41	0.0042	41.0	24994.2	209	0.0084	209.0	35988.2	556	0.0154	556.0

	70 to 74 Years				75 to 79 Years				80 to 84 Years			
Income	*E*	*O*	*M*	*EXP*	*E*	*O*	*M*	*EXP*	*E*	*O*	*M*	*EXP*
20,000–30,000	4198.0	157	0.0374	110.2	1870.9	112	0.0599	86.2	192.1	16	0.0833	13.8
30,000–40,000	5943.7	172	0.0289	156.1	2691.3	128	0.0476	124.1	341.0	22	0.0645	24.5
40,000–50,000	4706.3	94	0.0200	123.6	1913.5	95	0.0496	88.2	242.0	17	0.0702	17.4
50,000–60,000	3087.2	60	0.0194	88.2	1086.2	30	0.0276	50.1	181.2	16	0.0883	13.0
> 60,000	4193.9	98	0.0234	95.7	1875.6	70	0.0373	86.4	270.7	17	0.0628	19.4
Total	22129.1	581	0.0263	581.0	9437.5	435	0.0461	435.0	1227.0	88	0.0717	88.0

Notes: E = Exposed-to-risk; *O* = Observed number of deaths; *M* = Crude central mortality rates; and *EXP* = Expected deaths.

Although these data are available only for pensioners at quinquennial ages, Table 3 shows that the mortality rates calculated in this study are similar to the actuary's assumptions. This result is not surprising, as the actuary is likely to have based assumptions, at least in part, on the plan's experience.

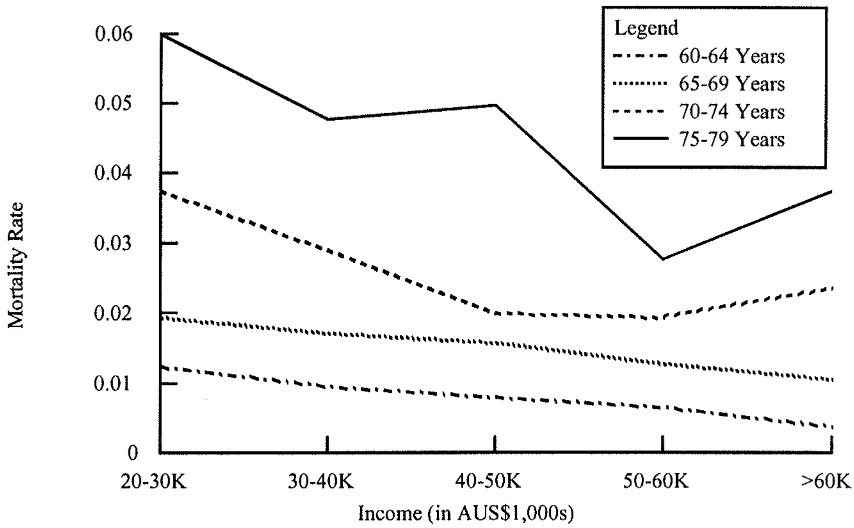
Table 3
Male Pensioner Mortality
Assumed vs. Actual Results

	Mortality Rates	
Age	Assumed	Actual
55	0.004	0.001
60	0.007	0.005
65	0.013	0.011
70	0.023	0.022
75	0.041	0.041
80	0.072	0.067

To test the hypothesis that mortality rates are equal at different income levels within a particular age group, the chi-square test statistic with 4 degrees of freedom is used. There is strong evidence to suggest that mortality rates are not equal at different income levels in the age groups 60 to 64 years ($p < 0.001$), 65 to 69 years ($p < 0.001$), 70 to 74 years ($p < 0.001$), and 75 to 79 years ($p < 0.001$). The chi-square statistic is not significant ($\alpha = 0.05$), however, for the 55 to 59 and 80 to 84 age groups. This lack of significance in these two groups is likely to be caused by different factors. First, the younger age group represents early retirements only, so this group is likely to have its own characteristics. Second, there are few data in the older age group. In addition, Wilkins, Adams, and Brancker (1989) show that the mortality disparities diminish markedly after age 74; it is possible that the lack of significance in the older age group also reflects a reducing effect at older ages. The likely presence of a selection effect will be discussed further in the next section.

Figures 1 and 2 display the observed mortality rates for the four age groups, where there are significant results, and highlight the general direction of the relationship between income and mortality rates. These graphs do not represent graduated curves but merely link the observed values.

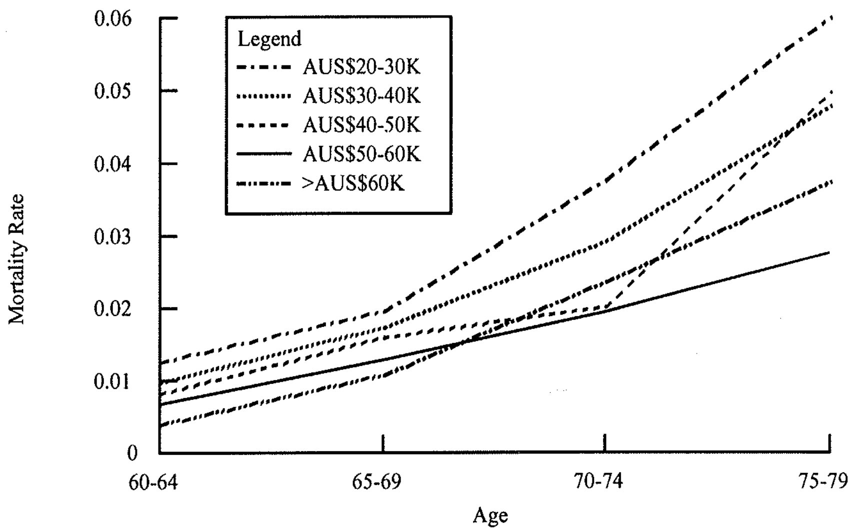
Figure 1
Mortality Rates and Income Levels



The data suggest that for male pensioners between the ages of 60 and 80, mortality is related to income with a trend toward lower mortality rates as income increases. This trend also can be seen from an inspection of the age-specific standardized mortality ratios (SMR) for each income group in Table 4. The standardized mortality ratio is the ratio of observed to expected deaths multiplied by 100.

As a measure of relative mortality between income levels we also can calculate the ratio of the SMR for the lowest income group to the SMR for the highest income group. Table 4 shows that the SMR of the lowest income level is over twice that of the highest income level for male pensioners ages 55 to 64. Although this disparity in death rates between high and low income earners is not as pronounced for the older age groups, the ratio of SMRs is always greater than one. These ratios also have been used in other international studies and therefore provide a useful point of comparison.

Figure 2
Mortality Rates by Age



3.2 International Comparisons

The direction of these Australian results agrees with other international studies, including a United States national longitudinal mortality study conducted under the auspices of the National Institutes of Health (Rogot, Sorlie, Johnson, and Schmitt, 1992). This 1979 to 1985 follow-up mortality study of 1.3 million persons involving twelve census sample cohorts found that white males age 55 or more exhibited an inverse relationship of mortality level with income. The standardized mortality ratio in the lowest income bracket was at least twice the standardized mortality ratio in the highest income bracket for white men age 55 to 64 years.

In Canada a collaborative study by Health and Welfare Canada and Statistics Canada (Wilkins, Adams, and Brancker, 1989) was conducted based on male residents of Canada's census metropolitan areas in 1986. Census tracts within each census metropolitan area were assigned to one of five income quintiles according to the proportion of residents with low total family income as determined by the national low income

Table 4
Age-Specific Standardized Mortality Ratios (SMR) by Income Level With the Income Range Expressed in 1994 Australian Dollars

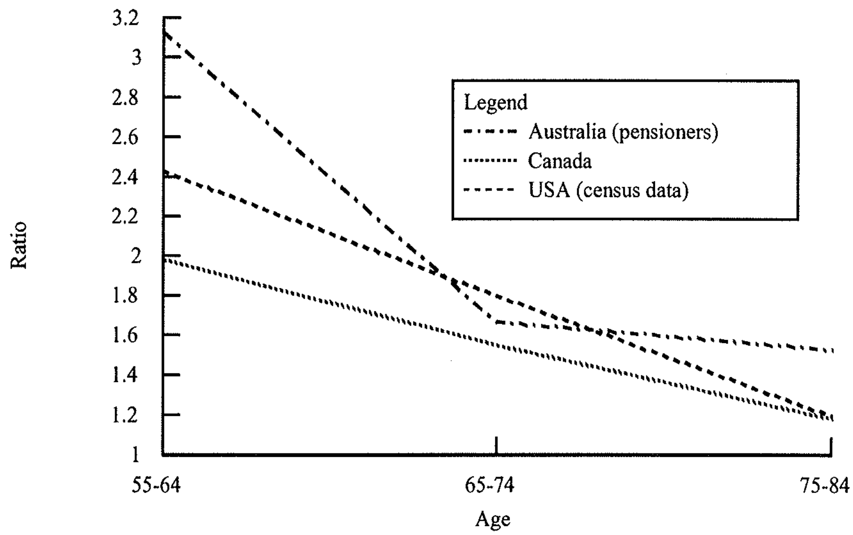
	Age Groups					
Income	55-59	60-64	65-69	70-74	75-79	80-84
20,000-30,000	131	148	126	142	130	116
30,000-40,000	94	115	111	110	103	90
40,000-50,000	116	94	103	76	108	98
50,000-60,000	101	80	83	74	60	123
> 60,000	49	47	68	89	81	88
Ratio of SMRs for:						
Min:Max Income	2.67	3.15	1.85	1.60	1.60	1.32
Extreme Values	2.67	3.15	1.85	1.92	2.17	1.32

cut-off. The Canadian data showed that for males ages 55 to 84 the higher the percentage of poor in a quintile, the higher the death rate.

Figure 3 shows the ratio of the age-specific SMR for the lowest income grouping to the age-specific SMR of the highest income group for each of these North American population studies as well as the Australian pensioner data for the same age groups. The similarity of the ratios, particularly for ages 65 to 84, is remarkable, given the different approaches taken. These three studies also confirm the suggestion that the income effect decreases with age.

A more directly comparable study with these Australian results is an investigation of the mortality of non-disabled annuitants in the United States covered under the Civil Service Retirement System (Virga, 1996). For the fiscal years 1988 to 1994 pensioners were pooled into five-year age groupings for each of five indexed final salary levels. Mortality rates declined significantly as the amount of final salary increased, with the differential between the highest and lowest salary levels also declining with increasing age. Figure 4 shows the ratio of the mortality rates of the annuitants in the lowest salary band (less than \$30,000 per annum) with the mortality rates for those in the highest salary band (greater than \$80,000 per annum) and compares them with the corresponding Australian age groups. The greater differential in mortality rates found for Australian pensioners age 55 to 64 years may reflect the fact that the definition of invalidity changed in June 1990, thereby making it harder for a person to receive an invalidity pension. Although it is conjectural,

Figure 3
Ratio of Age-Specific SMRs for Three Studies



it is possible that this change had a more significant impact on lower income earners who retired as normal pensioners in the early 1990s and were therefore under age 65 for the period under study.

The most recent study in the United Kingdom was presented in CMI Report Number 14 (1995) which notes that for pensioners covered by Life Office Pension Schemes: "It will once again be noted that the mortality recorded by reference to 'amounts' is significantly lighter than that recorded by reference to 'lives.' "

This result is consistent with the other studies, as amounts are likely to be related to income. The use of amounts is likely to be a less accurate proxy for socioeconomic factors than income, as a level of choice is present at the individual level. Furthermore, an annuity of a given amount could result from 15 years service for a final salary of X or 30 years service and a final salary of $0.5X$. In an environment where full portability of pensions does not exist and/or some individual choice is available, the use of final salary is likely to be a better indicator of the individual's socioeconomic position.

Other international studies have estimated the influence of wealth or total assets on mortality despite the difficulties involved in collecting comprehensive and accurate wealth data over a sufficient length of time (Attanasio and Hoynes, 1995; Menchik, 1993; Jianakoplos, Menchik, and Irvine, 1989). Each of these studies has found that an inverse relationship exists between wealth and mortality, thereby confirming the inverse relationship between mortality and financial well-being.

Furthermore, the studies from Australia, Canada, and the U.S.A. discussed above suggest a marked similarity in the experience of the three countries and a declining level of differential mortality as age increases. This latter result may be considered to be a form of a select period after retirement such that the significance of differential mortality reduces as the retiree ages.

The presence of a select period is further confirmed in the *U.K. CMI Report Number 14 (1995)*. Table 8.1.4 in this U.K. report shows that the ratio of the average pension of all exposed lives to the average pension of all deaths decreased steadily from 1.52 for the 61 to 65 age group to 1.23 for the 76 to 80 age group and then to 1.01 for the 91+ age group.

3.3 Female Pensioners

In this study the number of female pensioner records available is considerably fewer than the number of male pensioner records. Table 5 shows the experience and suggests that no consistent trend is evident across income levels for female pensioners. Test statistics are not significant at the 5 percent level for any age group.

While the evidence from various studies (Rogot et al., 1992; Virga, 1996; Wilkins, Adams, and Brancker, 1989) indicates that female mortality rates also vary inversely with income level, no such relationship is discernible from this, albeit smaller, set of data. One possible reason for the lack of a relationship between the level of income and rate of mortality for females is that many female pensioners worked during a period when the level of household income and/or wealth primarily was determined by the income earning capacity of the male member of the household.

It is also interesting to note that in the most directly comparable international study (Virga, 1996), there is evidence for differential mortality among female annuitants. The evidence is not as strong as the male ratios shown in Figure 4, with the female ratios for age groups 65 to 69 and 70 to 74 being 1.70 and 1.28, respectively, while the corresponding male ratios are 1.88 and 1.81.

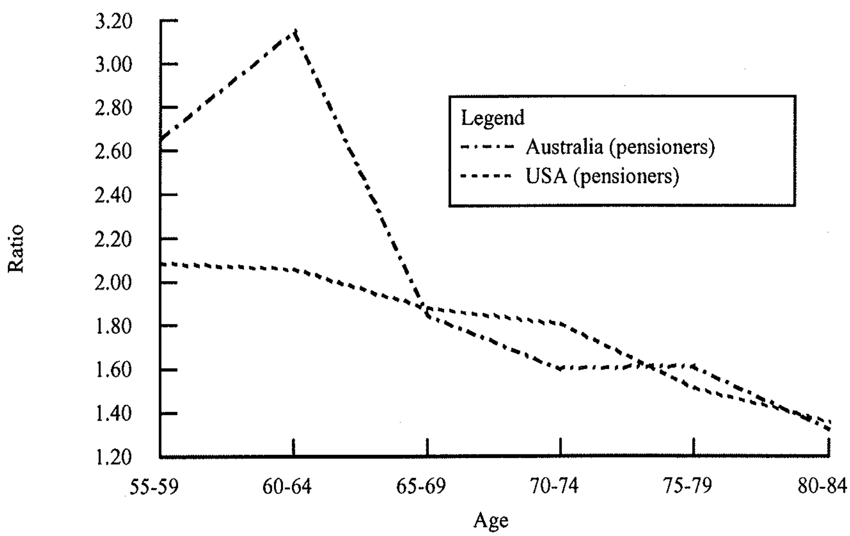
Table 5
Mortality Experience of Female Pensioners

	50 to 59 Years				60 to 64 Years				65 to 69 Years			
Income	*E*	*O*	*M*	*EXP*	*E*	*O*	*M*	*EXP*	*E*	*O*	*M*	*EXP*
20,000–30,000	1683.3	7	0.0042	6.0	4162.4	23	0.0055	24.5	4346.1	43	0.0099	42.8
30,000–40,000	1074.3	3	0.0028	3.9	2186.6	15	0.0069	12.9	2399.4	20	0.0083	23.6
> 40,000	584.9	2	0.0034	2.1	1125.3	6	0.0053	6.6	1171.8	15	0.0128	11.5
Total	3342.5	12	0.0036	12.0	7474.2	44	0.0059	44.0	7917.2	78	0.0099	78.0

	70 to 74 Years				75 to 79 Years				80 to 84 Years			
Income	*E*	*O*	*M*	*EXP*	*E*	*O*	*M*	*EXP*	*E*	*O*	*M*	*EXP*
20,000–30,000	2000.4	23	0.0115	21.4	550.1	15	0.0273	17.0	30.9	0	0.0000	1.1
30,000–40,000	1292.7	7	0.0054	13.8	458.2	20	0.0437	14.2	59.1	4	0.0677	2.1
> 40,000	732.0	13	0.0178	7.8	285.4	5	0.0175	8.8	24.4	0	0.0000	0.9
Total	4025.1	43	0.0107	43.0	1293.6	40	0.0309	40.0	114.4	4	0.0035	4.0

Notes: E = Exposed-to-risk; *O* = Observed number of deaths; *M* = Crude central mortality rates; and *EXP* = Expected deaths.

Figure 4
Mortality Ratio of Annuitants in Lowest Salary Band (< $30,000/Yr.) To Annuitants in Highest Salary Band (> $80,000/Yr.)



4 Some Consequences of Differential Mortality

4.1 Expectation of Life

One of the consequences of differential mortality is that there will be differing life expectancies for individuals of the same age. The extent of these differences will determine the significance of differential mortality for public pension policy, occupational pension funding, and the retail annuity market.

Assuming a uniform distribution of deaths for each year, the expectation of life at age x may be calculated as:

$$\mathring{e}_x = 0.5 + \sum_{n=1}^{\infty} \frac{l_{x+n}}{l_x}.$$

To determine the effect of differential mortality on life expectancy, the expectation of life is calculated for the following three groups of males pensioners in the study: (i) all male pensioners; (ii) male pensioners with a final salary between \$20,000 and \$30,000; and (iii) male pensioners with a final salary over \$60,000. Because the data for male pensioners over the age of 79 are limited, estimates of mortality rates at these older ages are based on the assumed mortality rates used in the government actuary's report. To find the corresponding age-specific rates for the two income groups, the ratio of the particular income group's mortality rate for ages 70 to 79 years to the corresponding figure for all income groups is determined and averaged over the 10 year age span. This differential ratio is reduced in a linear manner from age 75 to become 1.0 at age 100; this reducing ratio is used from age 80 onward. As a result of this process, mortality rates are calculated for all ages over 80 for the two income groups with the adjustments in the differential ratio allowing for the reducing effect of differential mortality with increasing age.

Table 6 shows the life expectancies for the total group and the two income groups. The disparity in mortality rates of male pensioners from different income groups means that, as shown in Figure 2, a 75 year old pensioner in the high income group has approximately the same risk of dying as a pensioner age 70 years from the low income group. Based on this study, a 65 year old male with a final salary of \$20,000 to \$30,000 has a life expectancy of 15.7 years, while those with a final salary of more than \$60,000 could be expected to live 18.9 years. While the differences in life expectancies at a particular age may not appear large, they may be significant for both the funding of occupational pensions and public policy considerations. These results also may be relevant with the growing consumer interest in pensions and a heightened concern for any systematic bias or inequity in these plans.

4.2 The Private (Voluntary) Annuity Market

Differentials in life expectancies can have significant implications for life insurance companies offering individual annuity products in a private sector market. The pricing of all insurance products (including annuities) takes into account a number of known, practical, and measurable factors that may influence the probability of a claim or, in the case of an annuity, the individual's life expectancy. Of course, it is also recognized that in some cases the insurer must ignore certain variables due to existing social custom, legislation, marketing pressure, or the difficulty in obtaining relevant data from the insured.

Table 6
A Comparison of Life Expectancies For Different Income Levels

	Income Levels		
Age	All	\$20K–\$30K	> \$60K
60	21.32	19.64	23.45
61	20.47	18.80	22.58
62	19.64	18.10	21.64
63	18.82	17.26	20.70
64	18.01	16.51	19.85
65	17.22	15.72	18.88
66	16.45	14.94	18.01
67	15.69	14.20	17.17
68	14.95	13.48	16.32
69	14.22	12.79	15.54
70	13.51	12.07	14.75
71	12.82	11.42	14.04
72	12.15	10.72	13.39
73	11.49	10.15	12.70
74	10.86	9.67	11.98
75	10.25	9.15	11.31
76	9.67	8.66	10.66
77	9.11	8.08	10.05
78	8.57	7.57	9.55
79	8.06	7.15	8.97
80	7.59	6.66	8.16

Notes: K denotes 1,000s.

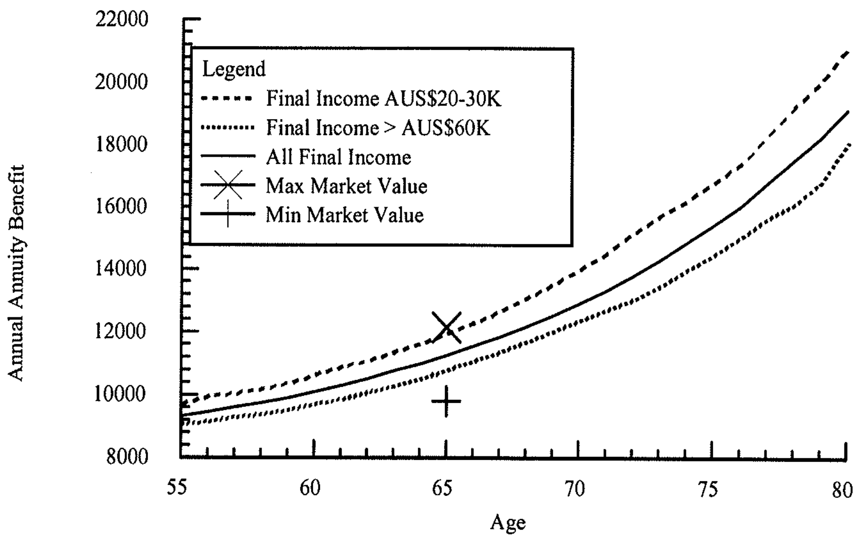
In terms of annuities for retirees, it may be particularly difficult or impractical to obtain appropriate information that enables the insurer to estimate the individual's socioeconomic status. Nevertheless, insurers are aware of differential mortality and the selection that occurs in the annuity market.

To further illustrate the effects of differential mortality, the level income stream arising from a life annuity with a present value of $100,000 is calculated corresponding to the life expectancies experienced by all male pensioners and by male pensioners in the highest and lowest income groups. An interest rate of 8 percent per annum is assumed and expenses are ignored. Figure 5 shows the level of annual income for this given purchase price for entry ages from 55 to 75. The calculated income levels are comparable with the immediate annuities currently offered by Australian life insurance companies. For example, a male age 65 years currently can purchase for $100,000 a level income (without any guarantee) between $9,804 and $12,136 per annum (Rice Kachor Research, 1996). If the annuity purchasers are primarily high income earners, the more aggressive life offices offering the higher income streams may be exposing themselves to a significant long-term risk. We also recognize that there are always many factors that affect market price.

Within many existing retail annuity markets the insurers assume that the expected mortality rates are equivalent for all individuals of the same age and gender irrespective of their lifetime income or accumulated wealth. Based on this and other studies, however, it is likely that lifetime income levels and/or wealth are important factors in determining life expectancies. It is therefore reasonable to expect that, on average, lifetime annuities will not be an attractive investment for individuals with lower incomes. On the other hand, higher income earners may find lifetime annuities, where mortality rates may be based on some population average together with some allowance for mortality improvement, to be an attractive proposition.

As was noted above, the pricing of annuities by an individual's lifetime income, final salary, or accumulated personal wealth would be an extremely difficult, if not impossible, task and could have a significant effect on the insurance company's reputation. Nevertheless, without some allowance for differential mortality in the pricing of annuities, it is reasonable to expect that the purchasers of annuities will experience below average mortality rates and that life insurance companies must allow for this adverse selection in their pricing processes.

Figure 5
Annual Level Lifetime Income Purchased for $100,000



4.3 Public Pension Policy

There exists an enormous variety of designs within public pension plans around the world. In some cases, a universal or means-tested age pension is paid from general taxation revenue with no direct link to an individual's taxation payments. That is, the size of the pension is not related to the individual's earning history. In other cases there is a relationship between the individual's pension contributions during his/her lifetime and the size of the pension received. In some cases a regressive scale exists such that the first tier of contributions results in a higher pension payment than subsequent contribution tiers. In other words, the rate of return received by the individual is higher for the band of contributions linked to lower incomes than for contributions related to higher salaries. Such a plan design is generally supported for reasons of intragenerational equity.

It is also important to consider the effects of differential mortality on the intragenerational equity within public plans. If higher income earners have longer life expectancies, then they will receive, on average, the public pension for a longer period of time. The actual effects will depend on the design of each program. If income redistribution is one objective of a public pension plan, then the achievement of this goal will be reduced and possibly reversed due to the existence of differential mortality.

For instance, in a flat rate universal pension program higher income earners will, on average, receive the age pension for an additional period (perhaps up to five years) than lower income earners. The possible inequity of this result needs to be appreciated. On the other hand, it is also important that one aspect of a particular plan should not be considered in isolation. That is, it may be necessary to review the effects in the context of the total taxation system for income and other retirement products. For instance, if higher income earners have paid considerably higher income taxes during their lifetime and/or pay higher taxes on the retirement benefits arising from their occupational and personal pension plans, it could be argued that the end result is not as inequitable as it may appear.

A different set of circumstances arises where the public pension is linked to lifetime earnings and/or contributions. In these cases, the higher income earner will be receiving a public pension that is both larger and is likely to be paid for a longer period than the lower income earner. Again, it is important to consider intragenerational equity issues within the context of all the issues including differential mortality, taxation, and government support.

Equity within public pension plans cannot be defined precisely and will vary according to the social and political decisions made by each community. Nevertheless, it is essential that the link between lifetime income and mortality is acknowledged and considered in the design of public plans. Furthermore, as the population gains a better understanding of public pension programs, it can be expected that systemic equity issues will be increasingly raised. It is therefore important that some data that measures the significance of differential mortality be available.

A related issue for a government's pension policy is linked to any legislation that may require the retirement benefit arising from an occupational pension plan to be taken, either wholly or partly, as a lifetime pension. The likely outcomes of such a policy are significant subsidies from low income earners to high income earners due to the differences in the expected longevity of each group. If such an outcome is considered undesirable, the effects of differential mortality could be ameliorated by an alternative pension arrangement. For instance, one possible solution is for the individual's pension to be paid from an allocated or segregated account, possibly with appropriate minimum and maximum limits, to ensure that the funds are preserved for a reasonable number of years. On early death the remaining assets could be passed to the individual's estate rather than used to support other pensioners. Such an arrangement would radically change the nature of a group pension plan, but it is consistent with recent developments in a number of countries and the growing importance of individual responsibility and entitlements. Such a development also reduces any intragenerational inequity that may arise due to differential mortality.

5 Conclusions

There is considerable international evidence suggesting that socioeconomic variables affect mortality rates. In practice, this means that there exists an inverse relationship between mortality rates and the level of lifetime earnings or wealth. The strength of this relationship has never been assessed among Australian retirees and has rarely been investigated for members of a single occupational pension plan.

This study, using data from the Australian Commonwealth Superannuation Scheme for public servants, shows that there is a significant inverse relationship for male pensioners between the individual's final salary and their rate of mortality in retirement. The results also confirm trends from previous North American and United Kingdom studies and suggest that there is a similar relationship between mortality in retirement and pre-retirement income in the United States, Canada, and Australia.

The strength of the relationship between an individual's income and mortality has important implications for the pricing of annuities in a voluntary private sector market. As income or wealth is not used in annuity pricing due to practical issues, it can be expected that an element of adverse selection will occur so that the mortality rates of annuitants would be considerably less than the population average. We also suggest that mortality assumptions used for the funding of occupational pensions should be adjusted to take into account the socioeconomic background of members.

The presence of differential mortality should be an important consideration in seeking intragenerational equity within public pension plans. The implications of pension policy will depend on a number of factors, including the design features of the public and occupational pension plans, the link between the size of any public pension and the level of lifetime contributions, the overall taxation structure, the strength of the differential mortality, and the social and political values of the society. Hence, there will be no one solution for all circumstances. An important outcome of these results is that policy makers recognize the existence and impact of differential mortality as they review the design and equity of public pension plans.

References

Atkinson, M.E., Creedy, J., and Knox, D.M. "Alternative Retirement Income Strategies: A Cohort Analysis of Lifetime Redistribution." *Economic Record* 72, 217 (June 1996): 97–106.

Attanasio, O.P., and Hoynes, H.W. "Differential Mortality and Wealth Accumulation" Working Paper No. 5126. Cambridge, Mass.: National Bureau of Economic Research, May, 1995.

Carney, T. and Hanks, P. *Social Security in Australia.* Melbourne, Australia: Oxford University Press, 1994.

Duval, D. *Public Sector Superannuation Scheme and Commonwealth Superannuation Scheme. A Report on Long-Term Costs.* Australian Government Publishing Service, 1994.

Institute of Actuaries and Faculty of Actuaries. *Continuous Mortality Investigation Report No. 14.* London: Institute of Actuaries and Faculty of Actuaries, (1995).

Jianakoplos, N.A., Menchik, P.L. and Irvine, F.O. "Using Panel Data to Assess the Bias in Cross-Sectional Inferences of Life-Cycle Changes in the Level and Composition of Household Wealth." In *Measurement of Savings, Investment and Wealth.* Chicago, Ill.: University of Chicago Press, 1989.

Menchik, P.L. "Economic Status as a Determinant of Mortality Among Black and White Older Males-Does Poverty Kill?" *Population Studies* 47, no. 3 (November 1993): 427–436.

Rice Kachor Research Pty Ltd. *Annuity and Pension League Table.* Sydney, Australia: Rice Kachor Research Pty Ltd, February 1996.

Rogot, E., Sorlie, P.D., Johnson, N.J. and Schmitt, C. *A Mortality Study of 1.3 Million Persons by Demographic, Social, and Economic Factors: 1979-1985 Follow-Up.* NIH Publication No. 92-3297. Bethesda, Md.: National Institutes of Health, National Heart, Lung and Blood Institute, 1992.

Virga, M.R. "Earn More, Live Longer-Variation in Mortality by Income Level." *Pension Section News* no. 28 (March 1996).

Wilkins, R., Adams, O. and Branicker, A. "Changes in Mortality by Income in Urban Canada from 1971 to 1986." *Health Reports 1/2* (1989): 137-174.

World Bank. *Averting the Old Age Crisis.* New York, N.Y.: Oxford University Press, 1994.

Journal of Actuarial Practice Vol. 6, 1998

Using Parametric Statistical Models to Estimate Mortality Structure: The Case of Taiwan

Shih-Chieh Chang *

Abstract†

A mixture parametric model is used to analyze the changing pattern of Taiwanese mortality from 1926 to 1991. Three different age ranges are modeled as mixtures of extreme value distributions, namely the Weibull, inverse Weibull, and Gompertz distributions. The results show a significant improvement of mortality over the years.

Key words and phrases: *mixture, extreme value distribution, Weibull, inverse-Weibull, and Gompertz*

1 Introduction

In a recent study of the mortality structure of the 1989 Taiwan Standard Ordinary Experienced Life Table, Chang (1995) observed that the mortality rates followed a parametric mixture model. Further, observing $\ln(q_x)$ from the published population life tables of Taiwan that were constructed between 1926 and 1991, a clear pattern emerges; see Figures 1 and 2.

---

*Shih-Chieh Chang, Ph.D., is an associate professor of risk management and insurance in the College of Commerce at National Chengchi University in Taiwan. Dr. Chang received his bachelor's degree in mathematics from National Taiwan University and his Ph.D. in statistics from the University of Wisconsin-Madison. His research interests are in stochastic models for actuarial science and the Taiwanese financial market data.

Dr. Chang's address is: Department of Risk Management and Insurance, College of Commerce, National Chengchi University, Taipei, Taiwan, R.O.C. Internet address: *bchang@nccu.edu.tw*

†The author wishes to thank the three anonymous referees and (especially) the editor for detailed comments and numerous suggestions on an earlier draft of this article.

Figure 1
Taiwanese Male Mortality Patterns From 1926 to 1991

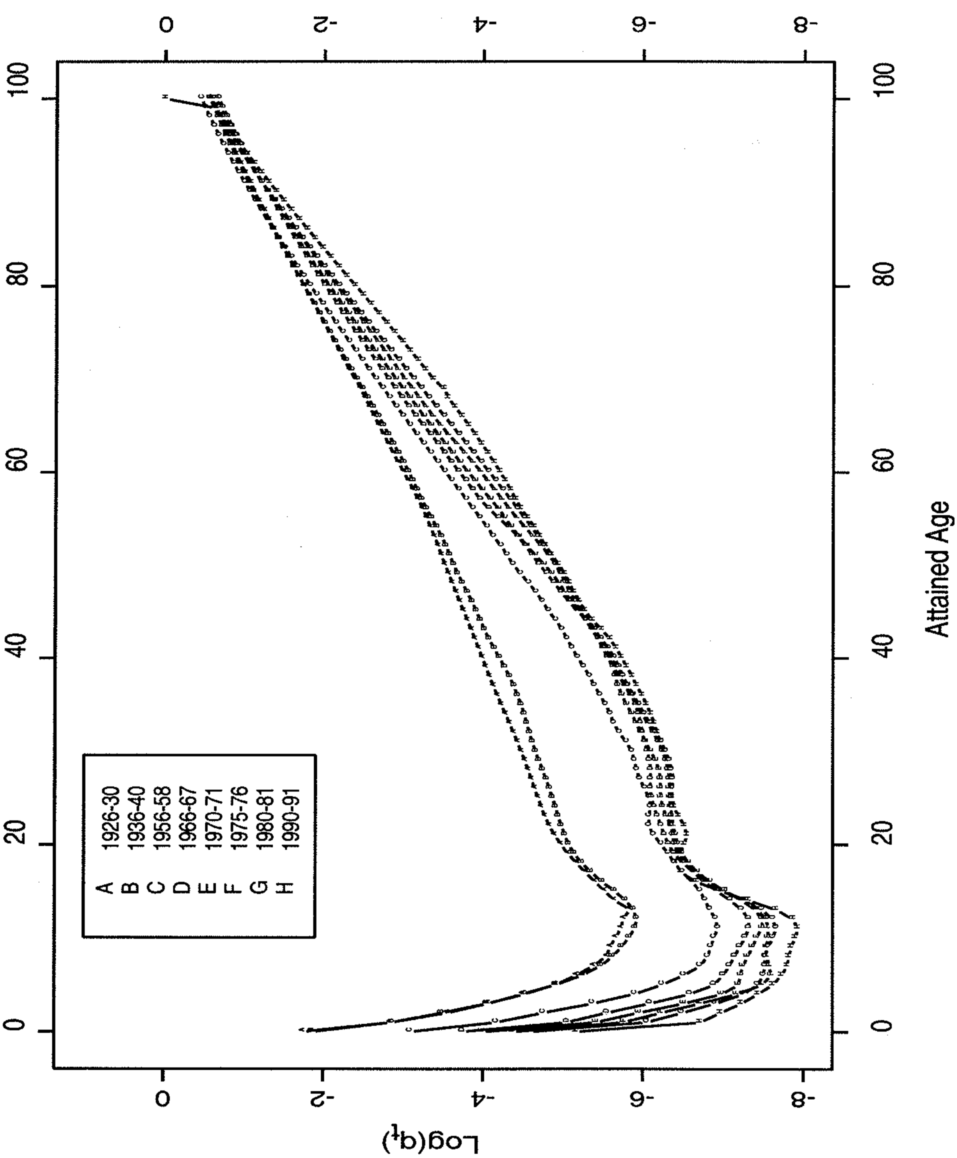
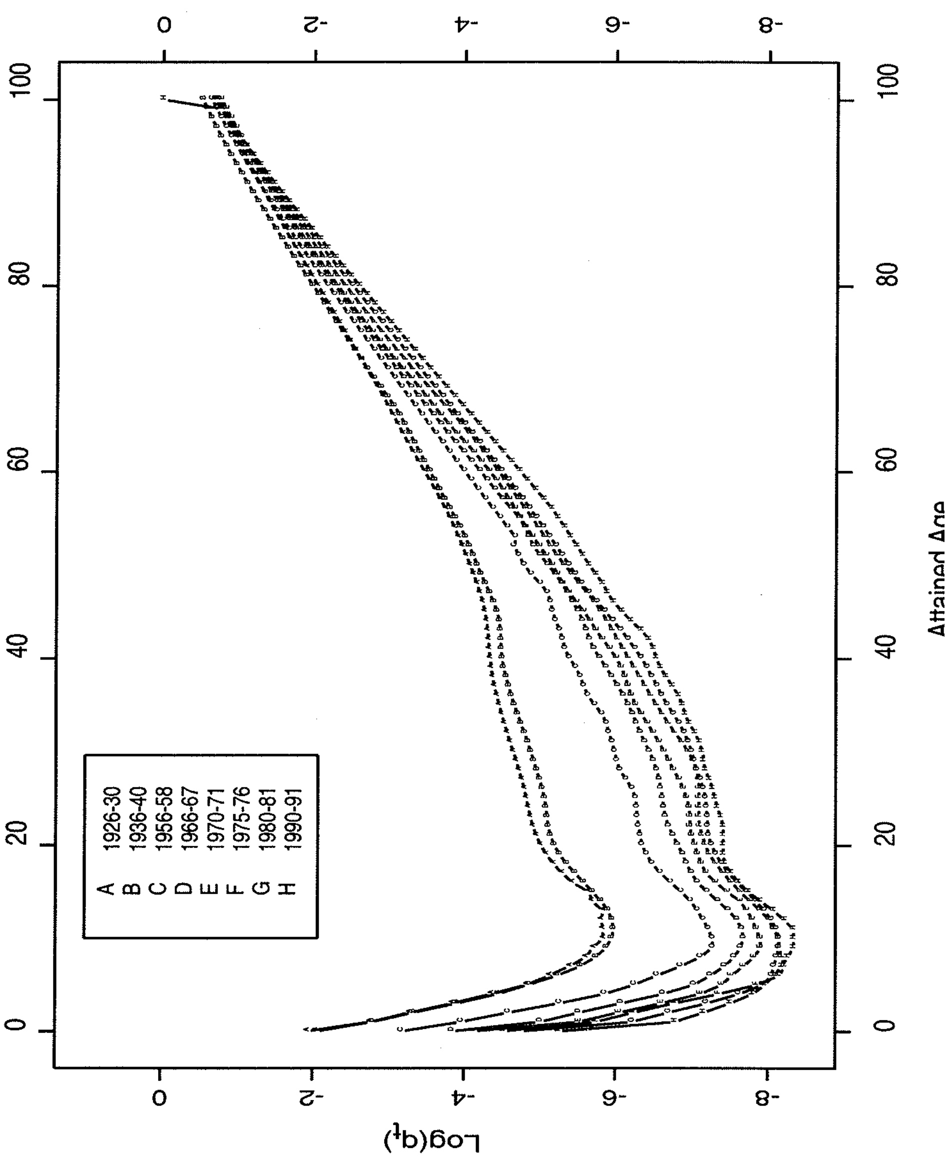


Figure 2
Taiwanese Female Mortality Patterns From 1926 to 1991



Figures 1 and 2 show similar mortality patterns for both male and female populations. The mortality rates appear to have three humps. These humps can be described by a parametric mixture model. The early life tables (constructed between 1926–1930 and 1936–1940) show a pattern of higher mortality rates. This was due, in part, to the effects of World War II and lack of adequate medical care in Taiwan's early years.

Parametric modeling is an important technique often used in the graduation of mortality rates; see, for example, Tenenbein and Vanderhoff (1980), Heligman and Pollard (1980), Wetterstrand (1981), Siler (1983), Renshaw (1991), Carriere (1992, 1994), Haberman and Renshaw (1996), and Yuen (1997). These authors have shown that parametric models provide an excellent means to understand a population's mortality structure.

A brief summary of advantages of using parametric models is listed below:

- Factors that influence (do not influence) mortality can be added (removed) from the model;
- If the results are not consistent with the proposed model, the model can be revised until it produces reasonable results;
- The resulting mortality rates form a smooth progression;
- Under certain loss criteria in the parameter estimation a general law of mortality can be obtained; and
- The final and most important advantage is the ability to forecast future mortality rates.[1]

2 Constructing the Parametric Model

The Taiwanese population mortality rates from 1926 to 1991 can be placed into three distinct subgroups that account for the major proportion of deaths:

- The infant population (ages 0–3);
- The adult population (ages 18–64); and
- The elderly population (ages 65 and over).

[1] A discussion of time trends, modeling, and forecasting can be found in Renshaw, Haberman, and Hatzopolous (1996).

Each of these population subgroups can be modeled by a distinct probability distribution. By combining these distributions, a finite mixture model can be used to analyze the entire population mortality rates.

Heligman and Pollard (1980) propose an eight-parameter model containing three distributions that fits Australian mortality rates. Carriere (1992 and 1994) uses a mixture of extreme value survival functions to model population mortality rates for a U.S. life table.

The parametric model is constructed as follows:

Step 1: Several well-known parametric statistical distributions, such as the Gompertz, Weibull, and inverse Weibull distributions, are chosen to see if they fit the mortality data. From Chang (1995), mixtures of these distributions have generated satisfactory results in estimating the Taiwanese life table.

Step 2: Simple graphical techniques are used to select the appropriate form of parametric distributions. For example, plots of $\ln(\hat{\mu}_x)$ vs x are examined, where $\hat{\mu}_x$ is the estimate of force of mortality at age x. If the plot appears to show a straight line, the Gompertz distribution might be appropriate to model the mortality data. If the pattern is shown to be a straight line in plots of $\ln(\hat{\mu}_x)$ vs $\ln(x)$ the Weibull distribution might be a better candidate. See Elandt-Johnson and Johnson (1980, Chapter 7) for more details on the use of such plots.

Step 3: Choose a base time point ($t = 0$) from which time is measured (in years). In this case we set January 1, 1926 as $t = 0$.

Step 4: For $t = 1, 2, \ldots$, let $s_t(x) \geq 0$ denote the survival function at t. We assume that $s_t(x)$ is a mixture of n component survival functions, i.e.,

$$s_t(x) = \sum_{i=1}^{n} \rho_{it} s_{it}(x) \quad \text{for } t = 1, 2, \ldots \tag{1}$$

where, for $i = 1, 2, \ldots, n$, $s_{it}(x) \geq 0$ is the ith component of the survival function, and $\rho_{it} \geq 0$ is the ith component mixing probability. Note that

$$\sum_{i=1}^{n} \rho_{it} = 1.$$

Step 5: The mixing probabilities and the parameters of each component $s_{it}(x)$ are estimated using statistical techniques.

For $i = 1, 2, \ldots, n$, let θ_{it} denote the vector used to describe the parameters in $s_{it}(x)$, i.e.,

$$s_{it}(x) = s_{it}(x \mid \theta_{it}). \tag{2}$$

Once the θ_{it}s are estimated, Carriere's (1994) select and ultimate parametric model is used to express θ_{it} as a function of t, i.e.,

$$\theta_{it} = \theta_{i0} + (\theta_{i\infty} - \theta_{i0})\left(1 - \exp(-a_i t^{b_i})\right), \quad a_i > 0,\ b_i > 0 \tag{3}$$

where a_i and b_i are parametric constants that influence the rate of convergence (as $t \to \infty$) of θ_{it} to $\theta_{i\infty}$. Equation (3) specifies the non-linear relationship between the respective elements of the vector of parameters and year.

Let $\theta_t = (\theta_{1t}, \ldots, \theta_{nt})$ and $\rho_t = (\rho_{1t}, \ldots, \rho_{nt})$ denote the vectors of parameters used to define $s_t(x)$. The proposed parametric model is:

$$s(x \mid \theta_t, \rho_t) = \sum_{i=1}^{n} \rho_{it} s_{it}(x \mid \theta_{it}), \tag{4}$$

which characterizes a general surface of the mortality rates.

Also, $s(x \mid \theta_t, \rho_t)$ can be regarded as a generalization of the model proposed by Wetterstrand (1981), who uses a Gompertz distribution (with year considered as a covariate) to analyze the mortality data.

3 The Life Tables Used and the Model

3.1 The Life Tables Used

The population life tables used are those published by the Taiwanese Department of Statistics, Ministries of Interior up to 1994. These tables are summarized in Table 1.

Though a life table may be released in a certain year, the table usually spans several years. For the purposes of this paper, a single year is assigned to each table. So, for the rth table releases since 1926, let t_r denote the value of t assigned to the table.

Table 1
Taiwanese Life Tables

Release Order (r)	t_r	Collection Period	Release Date
1	1926	1926–1930	1936, November
2	1935	1935–1940	1947, June
3	1956	1956–1958	1965, September
4	1966	1966–1967	1972, June
5	1970	1970–1971	1977, September
6	1975	1975–1976	1982, June
7	1980	1980–1981	1992, June
8	1990	1990–1991	1994, December

3.2 The Model

This study uses a model with three components ($n = 3$), each component being an extreme value distribution with a location parameter m_{it} and dispersion parameter σ_{it}. Thus $\theta_{it} = (m_{it}, \sigma_{it})$.

For the ith component at time t, the force of mortality $\mu_{it}(x)$, survival function $s_{it}(x)$, and probability density function $f_{it}(x)$ are summarized below:

Infants: The infant population (ages 0–3) is denoted by $i = 1$. This population is modeled as a Weibull distribution with parameters $m_{1t} > 0$ and $\sigma_{1t} > 0$: For $x \geq 0$,

$$
\begin{aligned}
\mu_{1t}(x) &= \frac{1}{\sigma_{1t}}\left(\frac{x}{m_{1t}}\right)^{\frac{m_{1t}}{\sigma_{1t}}-1}, \\
s_{1t}(x) &= \exp\left[-\left(\frac{x}{m_{1t}}\right)^{\frac{m_{1t}}{\sigma_{1t}}}\right], \\
f_{1t}(x) &= \frac{1}{\sigma_{1t}}\left(\frac{x}{m_{1t}}\right)^{\frac{m_{1t}}{\sigma_{1t}}-1}\exp\left\{-\left(\frac{x}{m_{1t}}\right)^{\frac{m_{1t}}{\sigma_{1t}}}\right\}.
\end{aligned}
$$

Adults: The adult population (ages 18–64) is denoted by $i = 2$. This population is modeled as an inverse Weibull distribution with parameters $m_{2t} > 0$ and $\sigma_{2t} > 0$: For $x \geq 0$,

$$\begin{aligned}
\mu_{2t}(x) &= \frac{\frac{1}{\sigma_{2t}}\left(\frac{x}{m_{2t}}\right)^{-\frac{m_{2t}}{\sigma_{2t}}-1}}{\exp\left(\left(\frac{x}{m_{2t}}\right)^{-\frac{m_{2t}}{\sigma_{2t}}}\right)-1}, \\
s_{2t}(x) &= 1-\exp\left(-\left(\frac{x}{m_{2t}}\right)^{-\frac{m_{2t}}{\sigma_{2t}}}\right), \\
f_{2t}(x) &= \frac{1}{\sigma_{2t}}\left(\frac{x}{m_{2t}}\right)^{-\frac{m_{2t}}{\sigma_{2t}}-1}\exp\left(-\left(\frac{x}{m_{2t}}\right)^{-\frac{m_{2t}}{\sigma_{2t}}}\right).
\end{aligned}$$

Elderly: The elderly population (ages 65 and over) is denoted by $i = 3$. This population is modeled as a Gompertz distribution with parameters $m_{3t} > 0$ and $\sigma_{3t} > 0$: For $x \geq 0$,

$$\begin{aligned}
\mu_{3t}(x) &= \frac{1}{\sigma_{3t}}\exp\left(\frac{x-m_{3t}}{\sigma_{3t}}\right), \\
s_{3t}(x) &= \exp\left(e^{-\frac{m_{3t}}{\sigma_{3t}}}-e^{\frac{x-m_{3t}}{\sigma_{3t}}}\right), \\
f_{3t}(x) &= \frac{1}{\sigma_{3t}}\exp\left(\frac{x-m_{3t}}{\sigma_{3t}}+e^{-\frac{m_{3t}}{\sigma_{3t}}}-e^{\frac{x-m_{3t}}{\sigma_{3t}}}\right).
\end{aligned}$$

Finally, let

$$\mathring{e}^{(it)}_{x:\overline{110-x}|} = \int_0^{110-x}\frac{s_{it}(x+y)}{s_{it}(x)}dy \tag{5}$$

denote the partial expectation of the future lifetime of a person between ages x and $110-x$ with survival function following $s_{it}(x)$. Because the tail probabilities of the Weibull distribution decrease to 0 slowly as $x \to \infty$, we choose the temporary complete life expectancy truncating at the limiting age at 110 instead of the complete future lifetime. For human lives, there have been few observations of age at death beyond 110; see, for example, Bowers et al., (1997, p. 86).

3.3 The Loss Function

In order to determine the parameters that best fit the data, a nonnegative loss function, $L_t(\theta_t, \rho_t)$, is used to measure adequacy of the estimation. The loss function is based on the sum of squared deviations.

The parameter estimates are determined by minimizing the combined loss function.

Let ω denote the highest age in the population life table; $q_{x,t}$ denote the observed mortality rate at age x and time t; $\hat{q}_{x,t}(\theta_t)$ denote the fitted mortality rate at age x and time t; and $w_{x,t}$ denote the weight assigned to age x and time t. In addition, let T denote the set of years assigned to the population life tables. From Table 1, the elements of T are $t_1, \dots, t_8$.

The loss function at time t, L_t and the combined loss function, L, are given by

$$L_t(\theta_t, \rho_t) = \sum_{x=0}^{\omega-1} w_{x,t}(\hat{q}_{x,t}(\theta_t) - q_{x,t})^2, \tag{6}$$

$$L = \sum_{t \in T} L_t(\theta_t, \rho_t). \tag{7}$$

Specifically, the weight function $w_{x,t} = 1/q_{x,t}^2$ is used in equation (6).[2] This weight function leads to the following loss function:

$$L_t(\theta_t, \rho_t) = \sum_{x=0}^{\omega-1} (1 - \frac{\hat{q}_{x,t}(\theta_t)}{q_{x,t}})^2 \quad \text{for } t \in T. \tag{8}$$

The minimization equation is:

$$\min L = \sum_{t \in T} L_t(\theta_t, \rho_t)$$

subject to $\sum_{i=1}^{3} \rho_{it} = 1$ for $t \in T$, $m_{it} > 0$, $\sigma_{it} > 0$ for $i = 1, 2, 3$ and $t \in T$. In the minimization process only the data between ages 0 and 90 are used because data over age 90 are difficult to obtain. Thus $\omega = 90$.

4 The Results

The computations in this paper were done using the software S-PLUS, which incorporates the S system developed at AT&T Bell Laboratories. Two basic references to S-PLUS are Becker, Chambers, and Wilks (1988) (for the programming aspects) and Chambers and Hastie (1992) (for the statistical modeling aspects).

[2]It is usual to use $w_{x,t} = 1/\sigma_{x,t}^2$ instead. So using $w_{x,t} = 1/q_{x,t}^2$ implies σ_x^2 is proportional to q_x^2, i.e., a constant coefficient of variation across age.

Table 2
Taiwan Male Population: Component 1
Estimates for Parameters and Partial Expectations of Life

r	Period	ρ_{1,t_r}	m_{1,t_r}	σ_{1,t_r}	$\mathring{e}^{(1,t_r)}_{0:\overline{110}\rceil}$	$\mathring{e}^{(1,t_r)}_{18:\overline{92}\rceil}$	$\mathring{e}^{(1,t_r)}_{65:\overline{45}\rceil}$
1	1926–30	0.2727	0.98	1.41	1.25	3.61	5.18
2	1936–40	0.2610	1.09	1.54	1.37	3.68	5.20
3	1956–58	0.0817	1.18	1.96	1.77	6.49	10.02
4	1966–67	0.0438	1.46	3.16	3.39	16.27	21.73
5	1970–71	0.0334	1.43	3.52	4.38	22.47	26.26
6	1975–76	0.0232	1.51	3.40	3.80	18.41	23.43
7	1980–81	0.0213	2.84	7.21	8.48	31.34	30.46
8	1990–91	0.0123	3.40	7.87	8.43	28.34	28.85

4.1 The Male Populations

Tables 2 to 4 present the parameter estimates and the partial expectations of life for the Taiwanese male population. The information in these tables is rearranged and displayed in Figures 3 through 8. In Table 2, the mixing probabilities $\rho_{1,t}$ decrease rapidly as t increases, while $m_{1,t}$ and $\sigma_{1,t}$ increase as t increases. Table 2 shows that the effect on the mortality rates from the infant population has diminished over time because $\rho_{1,t}$ decreases from 27.27 percent to 1.23 percent.

Table 3 shows that $\rho_{2,t}$ decreases from 3.54 percent to 2.16 percent gradually, while $m_{2,t}$ and $\sigma_{2,t}$ have shown no pattern over the years. In Table 4, $\rho_{3,t}$ shows a linear increasing trend from 69.19 percent to 96.61 percent over the years. In contrast, $m_{3,t}$ has increased from 61.87 to 79.86. There is no pattern for $\sigma_{3,t}$.

Figures 3 to 5 provide a better view of the parameter changes over the years. Figure 3 shows the mixture probabilities ($\rho_{i,t}$), Figure 4 shows the location parameters (m_{it}), and Figure 5 shows the dispersion parameters. Notice that the $\rho_{i,t}$ are decreasing by years in the infant population; are steady in the adult population; and, in the elderly population, increase to what appears to be their asymptotic values. In general, m_{it} and $\sigma_{i,t}$ have a tendency to increase both in the infant population and the elderly population, while $m_{i,t}$ remains level from about age 25 in the adult male population.

Table 3
Taiwan Male Population: Component 2
Estimates for Parameters and Partial Expectations of Life

r	Period	ρ_{2,t_r}	m_{2,t_r}	σ_{2,t_r}	$\mathring{e}^{(2,t_r)}_{0:\overline{110}\rceil}$	$\mathring{e}^{(2,t_r)}_{18:\overline{92}\rceil}$	$\mathring{e}^{(2,t_r)}_{65:\overline{45}\rceil}$
1	1926-30	0.0354	25.20	6.61	31.15	13.55	17.96
2	1936-40	0.0322	23.47	6.43	29.39	12.37	18.54
3	1956-58	0.0275	25.96	12.19	38.97	23.96	26.44
4	1966-67	0.0284	25.47	9.63	35.31	19.02	23.27
5	1970-71	0.0280	27.13	11.34	38.80	22.50	24.82
6	1975-76	0.0294	26.41	11.64	38.62	22.94	25.55
7	1980-81	0.0234	24.45	8.64	33.16	16.90	22.24
8	1990-91	0.0216	23.58	8.01	31.59	15.49	21.64

Table 4
Taiwan Male Population: Component 3
Estimates for Parameters and Partial Expectations of Life

r	Period	ρ_{3,t_r}	m_{3,t_r}	σ_{3,t_r}	$\mathring{e}^{(3,t_r)}_{0:\overline{110}\rceil}$	$\mathring{e}^{(3,t_r)}_{18:\overline{92}\rceil}$	$\mathring{e}^{(3,t_r)}_{65:\overline{45}\rceil}$
1	1926-30	0.6919	61.87	16.60	53.96	38.06	8.69
2	1936-40	0.7068	63.93	15.61	56.10	39.77	8.88
3	1956-58	0.8908	72.59	11.31	66.18	48.53	10.30
4	1966-67	0.9278	75.61	10.68	69.51	51.72	11.67
5	1970-71	0.9386	76.14	10.83	70.22	52.44	12.23
6	1975-76	0.9474	77.33	10.82	71.14	53.35	12.79
7	1980-81	0.9553	77.98	10.99	71.70	53.91	13.27
8	1990-91	0.9661	79.86	11.49	73.30	55.53	14.70

Figure 3
Mixture Probabilities for Taiwanese Male Tables 1926-1991

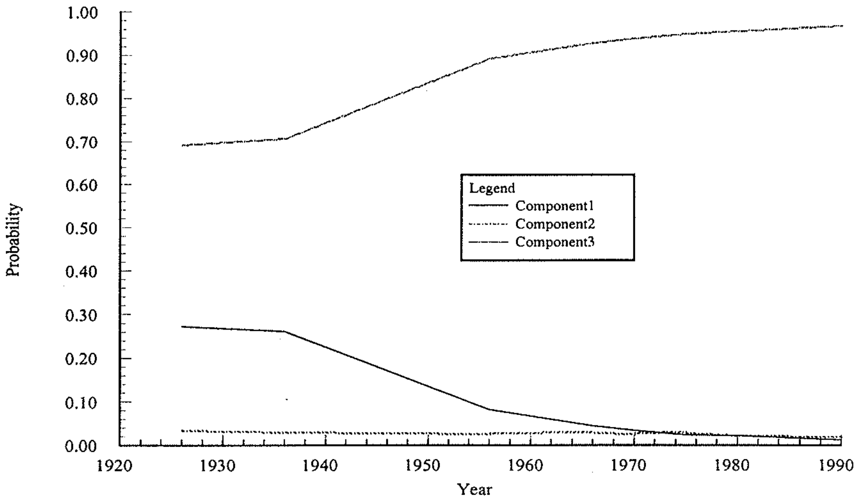


Figure 4
Location Parameters for Taiwanese Male Tables 1926-1991

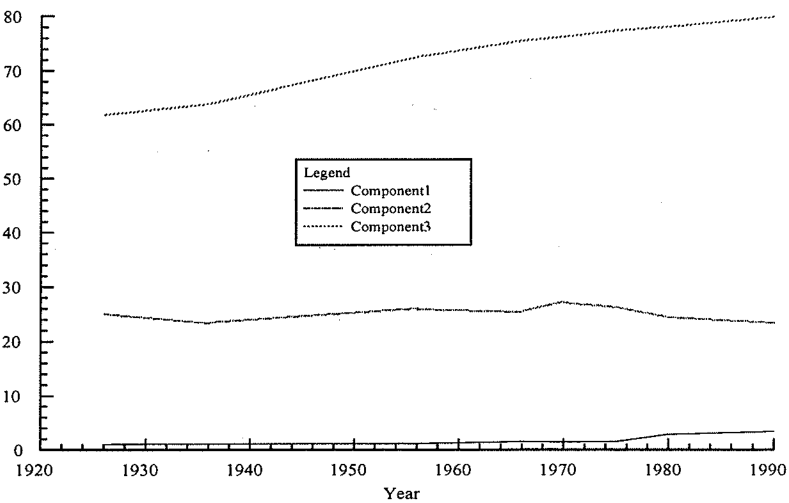
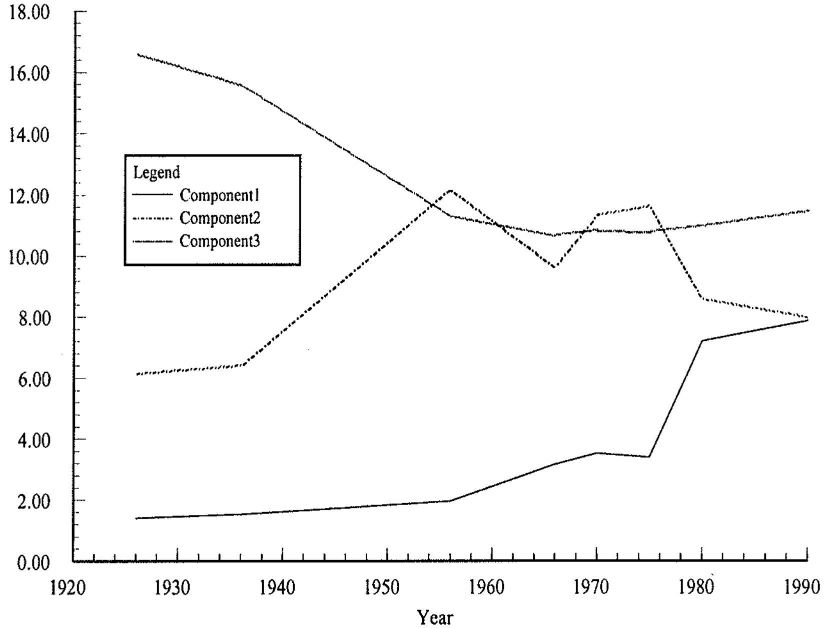


Figure 5
Dispersion Parameters for Taiwanese Male Tables 1926–1991



The partial expectations of life for the male population in Table 2 to Table 4 have shown increasing trends over the years both in component 1 and component 3, while the same pattern does not appear in component 2. Table 4 displays that the partial expectations of life in component 3 increase gradually from 53.96 to 73.30 at age 0; 38.06 to 55.33 at age 18; and 8.69 to 14.70 at age 65 .

Figures 6 through 8 compare the partial expectations of life over the years with each component. In component 1, it has shown an increasing trend indicating that the future lifetime is increasing at different ages. This increasing pattern in component 3 has explained the aging pattern in Taiwan, while the pattern in component 2 is not clearly shown.

Figure 6
$\mathring{e}_{0:\overline{110}|}$ for Taiwanese Male Tables 1926–1991

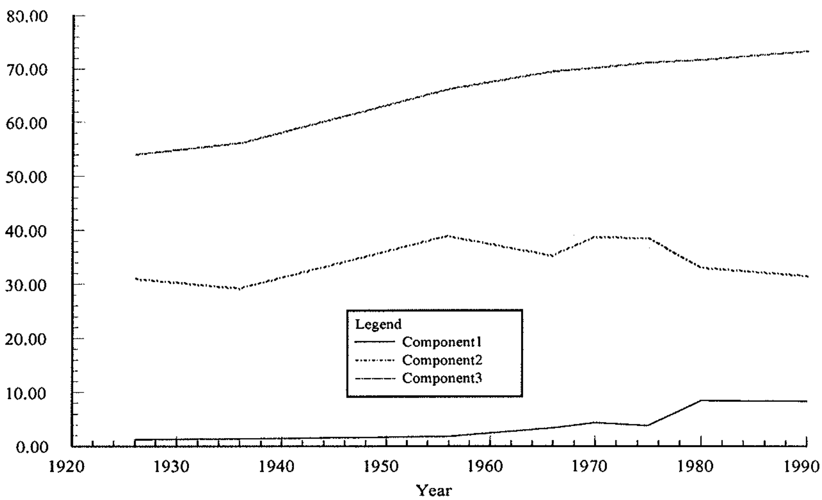


Figure 7
$\mathring{e}_{18:\overline{92}|}$ for Taiwanese Male Tables 1926–1991

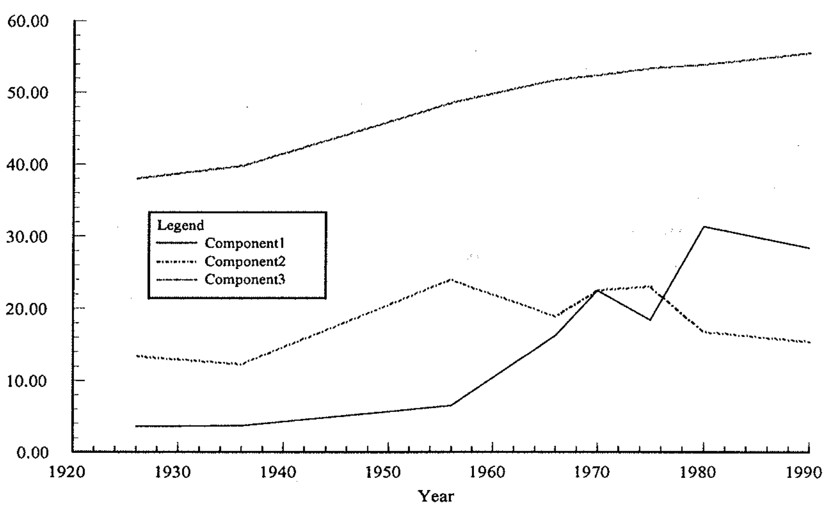
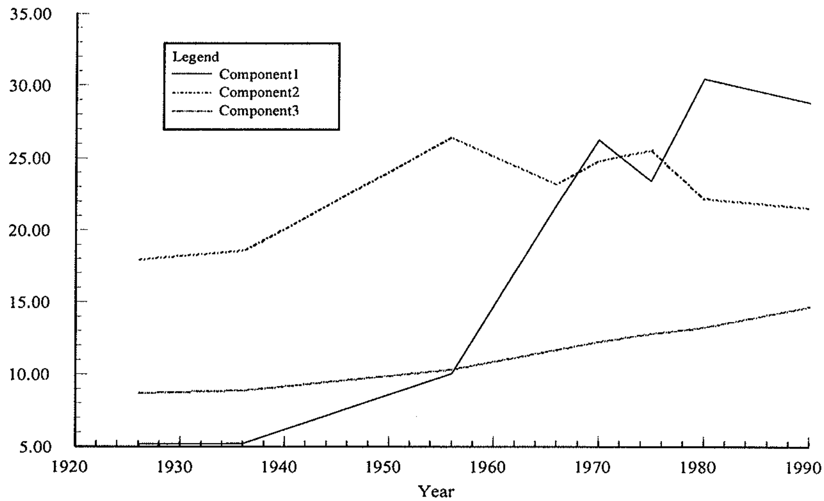


Figure 8
$\mathring{e}_{65:\overline{45|}}$ for Taiwanese Male Tables 1926–1991



4.2 The Female Populations

The parameter estimates for Taiwanese females are listed in Tables 5 to 7 and displayed in Figures 9 through 14.

The results for females are largely similar to those for males, especially for the infant and elderly populations. In contrast to the results obtained for the males, however, $m_{2,t}$ and $\sigma_{2,t}$ for the female populations are more unstable. Notice that $m_{2,t}$ has increased in the 1975–1976 table, then decreased in the 1990–1991, and $\sigma_{2,t}$ shows the same pattern; (see Figures 9 to 11).[3]

Like the male population, the partial expectations of life for the female population in Tables 5 to 7 also show an increasing trend over the years in components 1 and 3 and, to a lesser extent, in component 2.

4.3 Parameter Asymptotics

Notice that the population life tables in 1926–1930 and 1936–1940, which were constructed before and during World War II, yield signif-

[3]This pattern may be due to the change in the socioeconomic status of women. Further analysis is needed to explore this.

Table 5
Taiwan Female Population: Component 1
Estimates for Parameters and Partial Expectations of Life

r	Period	ρ_{1,t_r}	m_{1,t_r}	σ_{1,t_r}	$\mathring{e}^{(1,t_r)}_{0:\overline{110}\vert}$	$\mathring{e}^{(1,t_r)}_{18:\overline{92}\vert}$	$\mathring{e}^{(1,t_r)}_{65:\overline{45}\vert}$
1	1926-30	0.2756	1.24	1.97	1.75	5.85	8.85
2	1936-40	0.2549	1.33	1.99	1.77	5.15	7.45
3	1956-58	0.0817	1.35	1.84	1.64	3.85	5.27
4	1966-67	0.0388	1.29	2.04	1.82	5.93	8.93
5	1970-71	0.0272	1.14	2.12	2.01	9.02	14.01
6	1975-76	0.0186	1.22	1.71	1.52	3.89	5.46
7	1980-81	0.0153	1.36	2.85	2.97	14.29	20.01
8	1990-91	0.0124	4.59	11.57	12.16	36.38	32.31

icantly higher mortality than the other tables for both males and females. So, to determine the parameters that fit equation (3), we only consider the estimates from the life tables after 1940 (to obtain more consistent results). The parameters of equation (3) are estimated using *NLMIN*, the nonlinear optimization procedure in S-PLUS.[4] The estimates and asymptotes (as $t \to \infty$) are given in Table 8.

The Kolmogorov-Smirnov goodness-of-fit test is used with the null hypothesis that the distributions of these two samples are the same.[5] The critical value for D_n with a 5 percent significance level is approximately $1.36/\sqrt{90} = 0.14335$. The D_ns for the fitted model are summarized in Table 9. Because the D_ns are less than 0.14335, the results again support the use of the mixture parametric model to fit the population mortality rates.

[4]*NLMIN* is based on a quasi-Newton method using double dogleg step with BFGS secant update to the Hessian. For more details, see Dennis, Gay, and Welsch (1981) and Dennis and Mei (1979).

[5]The Kolmogorov-Smirnov goodness-of-fit statistic (D_n) is defined as

$$D_n = \sup_{0 \le x \le 89} [\mid s(x) - s_n(x) \mid]$$

where n is 90. See Hogg and Klugman (1984) or Hogg and Tanis (1983) for a detailed discussion on this statistic.

Table 6
Taiwan Female Population: Component 2
Estimates for Parameters and Partial Expectations of Life

r	Period	ρ_{2,t_r}	m_{2,t_r}	σ_{2,t_r}	$\overset{\circ}{e}{}^{(2,t_r)}_{0:\overline{110}\rceil}$	$\overset{\circ}{e}{}^{(2,t_r)}_{18:\overline{92}\rceil}$	$\overset{\circ}{e}{}^{(2,t_r)}_{65:\overline{45}\rceil}$
1	1926-30	0.1068	27.08	10.61	37.85	21.20	23.87
2	1936-40	0.0914	26.58	11.12	38.10	22.00	24.80
3	1956-58	0.0582	33.49	21.39	52.89	37.78	31.14
4	1966-67	0.0537	39.44	29.48	60.94	45.75	33.46
5	1970-71	0.0385	40.34	30.74	42.87	24.87	6.61
6	1975-76	0.0376	43.48	38.93	66.01	51.84	35.60
7	1980-81	0.0211	33.31	22.43	53.45	38.93	31.79
8	1990-91	0.0197	32.80	19.94	51.39	36.04	30.46

Table 7
Taiwan Female Population: Component 3
Estimates for Parameters and Partial Expectations of Life

r	Period	ρ_{3,t_r}	m_{3,t_r}	σ_{3,t_r}	$\overset{\circ}{e}{}^{(3,t_r)}_{0:\overline{110}\rceil}$	$\overset{\circ}{e}{}^{(3,t_r)}_{18:\overline{92}\rceil}$	$\overset{\circ}{e}{}^{(3,t_r)}_{65:\overline{45}\rceil}$
1	1926-30	0.6176	70.62	14.49	62.84	45.83	11.12
2	1936-40	0.6537	71.71	13.52	64.29	47.03	11.09
3	1956-58	0.8601	77.38	10.89	71.16	53.36	12.85
4	1966-67	0.9075	79.14	10.04	73.37	55.49	13.61
5	1970-71	0.9343	79.92	10.07	74.13	56.24	14.14
6	1975-76	0.9454	81.17	9.78	75.54	57.63	14.90
7	1980-81	0.9659	82.14	9.92	76.43	58.52	15.64
8	1990-91	0.9679	83.57	9.40	78.15	60.20	16.54

Table 8
Parameters from the Nonlinear Estimation

Parameter	1956-58	Asymptote	a	b
$\rho_{3,t}$(%) (female)	77.38	85.64	0.0072	1.50
$m_{3,t}$ (female)	86.01	97.02	0.0115	1.69
$m_{3,t}$ (male)	89.07	99.10	0.0571	0.91

Table 9
Kolmogorov-Smirnov Statistic D_n

Year	Male	Female
1926-30	0.00835	0.012511
1936-40	0.00272	0.012900
1956-58	0.00139	0.006344
1966-67	0.01995	0.002678
1970-71	0.00732	0.004283
1975-76	0.00204	0.004531
1980-81	0.00618	0.003248
1990-91	0.01338	0.005294

5 Closing Comments

Extreme value distributions are the underlying distributions of the parametric mixture model used to analyze the mortality structure in Taiwan from 1926 to 1991. This approach provides a more detailed model to understand the changing mortality pattern over the years and may form a better basis for projecting future mortality rates.

The mixture model points out the different patterns of mortality between Taiwanese male and female populations in different age components. The gender differences may be explained by the different socioeconomic roles men and women play in Taiwanese society.

Further research is needed to apply these results to premium and reserve calculations and to construct appropriate parametric models that include the effects of specific demographic impacts on mortality.

Figure 9
Mixture Probabilities for Taiwanese Female Tables 1926–1991

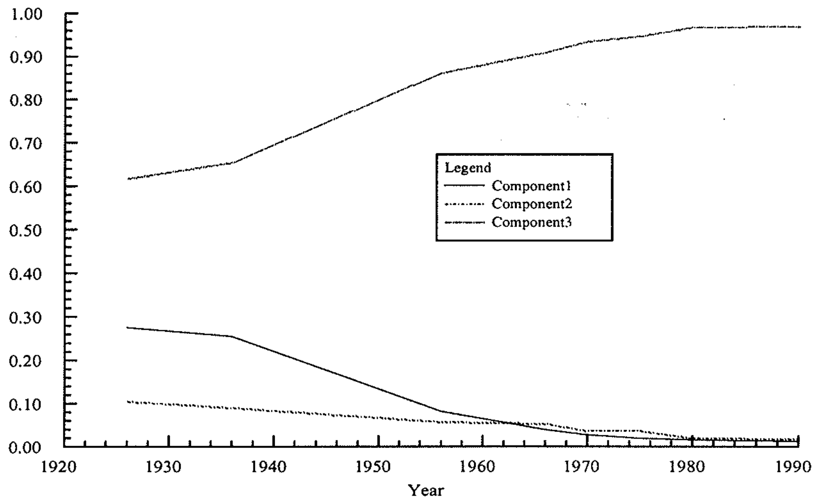


Figure 10
Location Parameters for Taiwanese Female Tables 1926–1991

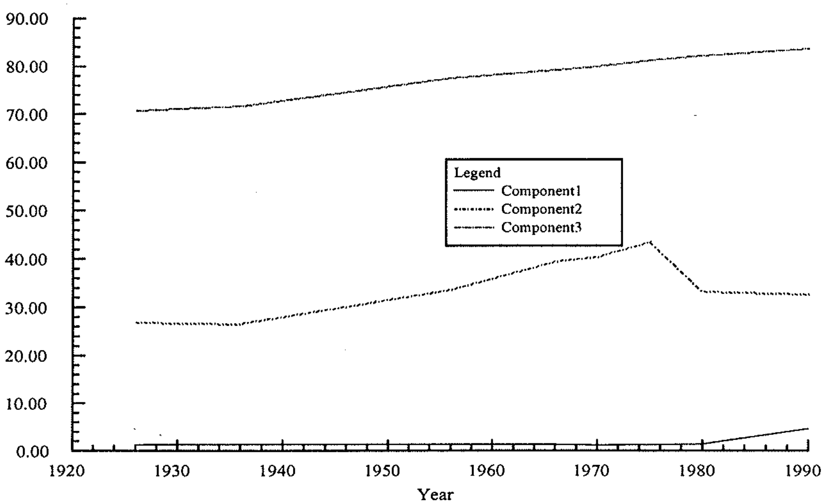


Figure 11
Dispersion Parameters for Taiwanese Female Tables 1926–1991

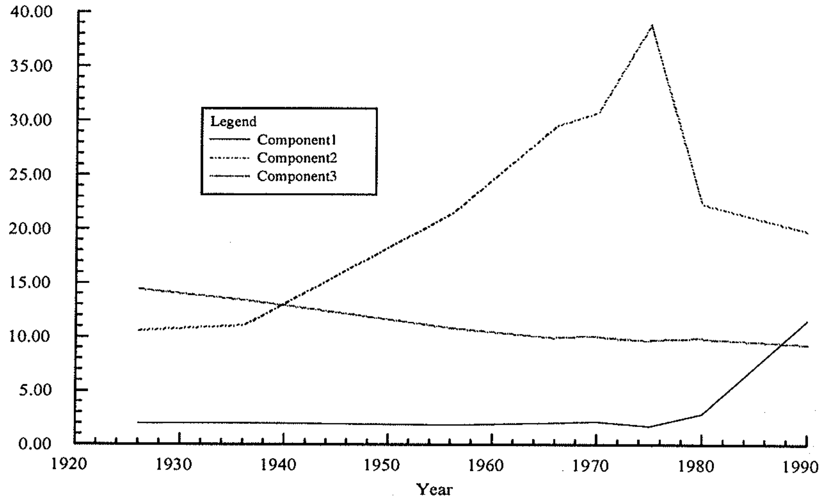


Figure 12
$\mathring{e}_{0:\overline{110}|}$ for Taiwanese Female Tables 1926–1991

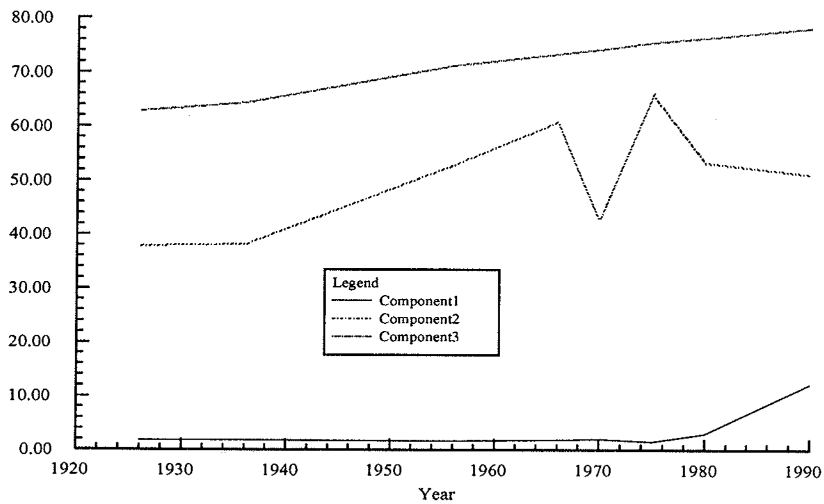


Figure 13
$\mathring{e}_{18:\overline{92}|}$ for Taiwanese Female Tables 1926–1991

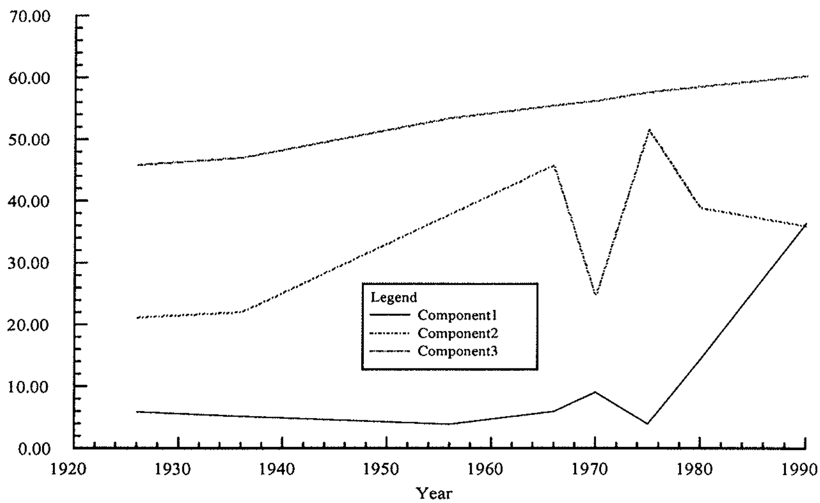
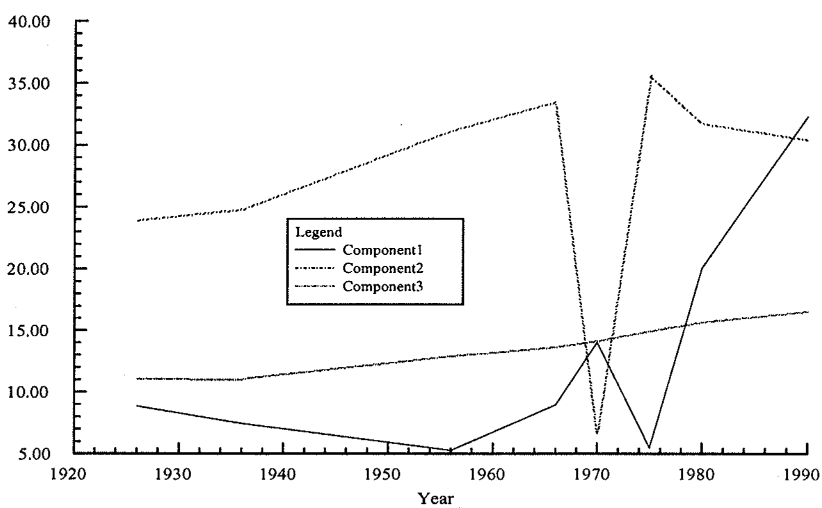


Figure 14
$\mathring{e}_{65:\overline{45}|}$ for Taiwanese Female Tables 1926–1991



References

Becker, R.A., Chambers, J.M. and Wilks, A.R. *The New S Language.* New York, N.Y.: Chapman and Hall, 1988.

Bowers, N.L., Jr., Gerber, H.U., Hickman, J.C., Jones, D.A. and Nesbitt, C.J. *Actuarial Mathematics, 2nd ed.* Schaumburg, Ill.: Society of Actuaries, 1997.

Carriere, J.F. "Parametric Models for Life Tables." *Transactions of the Society of Actuaries.* 44 (1992): 77-99.

Carriere, J.F. "A Select and Ultimate Parametric Model." *Transactions of the Society of Actuaries.* 46 (1994): 75-93.

Chambers, J.M. and Hastie, T.J. (eds.) *Statistical Models in S.* New York, N.Y.: Chapman and Hall, 1992.

Chang, S.C. "Using Mixture Parametric Model to Analyze the 1989 Taiwan Standard Ordinary Life Table." *Insurance Monograph (in Chinese)* 42 (1995): 112-118.

Dennis, J.E., Gay, D.M. and Welsch, R.E. "An Adaptive Nonlinear Least-Squares Algorithm." *ACM Transactions on Mathematical Software* 7 (1981): 348-383.

Dennis, J.E. and Mei, H.H.W. "Two New Unconstrained Optimization Algorithms Which Use Function and Gradient Values." *Journal of Optimization Theory and Applications.* 28 (1979): 453-483.

Elandt-Johnson, R.C. and Johnson N.L. *Survival Models and Data Analysis.* New York, N.Y.: John Wiley & Son Inc., 1980.

Haberman, S. and Renshaw, A.E. "Generalized Linear Models and Actuarial Science." *The Statistician* 45 (1996): 407-436.

Heligman, L.M.A. and Pollard, J.H. "The Age Pattern of Mortality." *Journal of the Institute of Actuaries* 107, part 1 (1980): 49-82.

Hogg, R.V. and Klugman, S.A. *Loss Distributions.* New York, N.Y.: John Wiley & Sons Publications, Inc., 1984.

Hogg, R.V. and Tanis, E.A. *Probability and Statistical Inference, 2nd ed.* New York N.Y.: Macmillian Publications, Inc., 1983.

Renshaw, A.E. "Actuarial Graduation Practice and Generalized Linear and Nonlinear Models." *Journal of Institute of Actuaries* 118 (1991): 295-312.

Renshaw, A.E., Haberman, S. and Hatzopolous, P. "Recent Mortality Trends in UK Male Assured Lives." *British Actuarial Journal* 2 (1996): 449-477.

Siler, W. "Parameters of Mortality in Human Populations with Widely Varying Life Spans." *Statistics in Medicine* 2 (1983): 373-380.

Tenenbein, A. and Vanderhoof, I. "New Mathematics Laws of Select and Ultimate Mortality." *Transactions of the Society of Actuaries* 32 (1980): 119—158.

Wetterstrand, W.H. "Parametric Models for Life Insurance Mortality Data: Gompertz's Law Over Time." *Transactions of the Society of Actuaries* 33 (1981): 159-179.

Yuen, K.C. "Comments on Some Parametric Models for Mortality Tables." *Journal of Actuarial Practice* 5, no. 2 (1997): 253-266.

A Frailty Model for Projection of Human Mortality Improvements

Shaun S. Wang* and Robert L. Brown†

Abstract‡

Based on the everyday observations that individual human beings vary significantly in their capacity to combat death, we adopt a so-called frailty model of human mortality. This frailty model assumes that each individual in a given population is endowed with his or her own frailty index, r, which remains constant for life. In addition, we assume that the individual's force of mortality (hazard rate function) at age x, $\mu_x(r)$, satisfies $\mu_x(r) = r\mu_x$ where μ_x is the population's base force of mortality at age x. Given the probability distribution of the frailty index among the newborns in the population, an expression is given for the distribution of the frailty index among the survivors reaching age

*Shaun S. Wang, Ph.D., A.S.A., is a pricing actuary at the SCOR Reinsurance Company. He also is an adjunct associate professor of actuarial science at the University of Waterloo. Dr. Wang has published numerous articles in actuarial journals including *Astin Bulletin, Insurance: Mathematics and Economics, North American Actuarial Journal,* and *Proceedings of the Casualty Actuarial Society,* His 1997 paper entitled: "Implementation of PH-Transform in Ratemaking" won the best paper award for the Casualty Actuarial Society Ratemaking Call Paper Program.

Dr. Wang's address is: SCOR RE, One Pierce Place, PO Box 4049, Itasca IL 60143, U.S.A. Internet address: *swang@scor.com*

†Robert L. Brown, Ph.D., F.S.A., F.C.I.A., A.C.A.S., is a professor of actuarial science and director of the Institute of Insurance and Pension Research at the University of Waterloo, Canada. Dr. Brown was president of the Canadian Institute of Actuaries in 1990/91 and served on the Board of Governors of the Society of Actuaries from 1992 to 1996. In 1997 he was elected vice president of the Society of Actuaries. He was a city councillor in Waterloo from 1988 to 1994. Dr. Brown has authored five books and several papers. In 1994 he won the S.C.O.R. International Papers Competition for his paper: "Pay-As-You-Go Funding Stability and Intergenerational Equity."

Dr. Brown's address is: Department of Statistics and Actuarial Science, University of Waterloo, Waterloo ON N2l 3G1, CANADA. Internet address: *rl-brown@jeeves.uwaterloo.ca*

‡The authors wish to thank Jiahua Chan, Harry Panjer, and Gordon Willmot of the University of Waterloo; Virginia Young of the University of Wisconsin; and the referees and the editor for numerous valuable suggestions and comments.

x in the population. Finally, assuming that (i) the rate of mortality improvement for any age is proportional to the average frailty level of the individuals at that age, (ii) a gamma distribution for the frailty index, and (iii) a Gompertz form for the population's base force of mortality, we graduate (smooth) the observed mortality improvement factors in the published Society of Actuaries' GAR-94 Table.

Key words and phrases: *force of mortality, hazard rate, gamma distribution, Gompertz law*

1 A Review of Actuarial Mortality Projection

Throughout most of the twentieth century (except during periods of famine, war, and other civil strife), there has been a long and consistent trend of mortality improvement. Lancaster (1990, Chapter 3.6, Table 3.6.1) shows the persistent decline in the overall mortality in several western countries. The reason for this decline is largely because of improvements in public health, improvements in the production and distribution of food, and advances in medicine and technology.

Interestingly, Vaupel and Yashin (1987, pp. 123) note that progress in reducing mortality can be conceived in two ways. Demographers generally view mortality change as change in the force of mortality and associated life table statistics for a population. Most relative laypersons, on the other hand, especially physicians and other health and safety personnel, perceive a reduction in mortality as being achieved by saving the lives of individuals faced with death. A demographer might report that the force of mortality at age fifty among U.S. males was cut in half from 1900 to 1980, from 1.6 percent to 0.8 percent. A public health specialist might focus attention on the lives that were saved in 1980 compared with 1900 because of new surgical and medical procedures, the introduction of penicillin, polio vaccines, and other pharmaceuticals, better nutrition and sanitation, improved automotive safety, a decrease in cigarette smoking, faster and more effective ambulance service, and so on.

Actuaries, like demographers, generally view mortality change as change in the force of mortality and associated life table statistics for a population. In fact, the projection of mortality improvement has been an important subject to actuaries. For example, in the first issue of the *Transactions of the Society of Actuaries* Jenkins and Lew (1949) give a lengthy discussion on this subject. Over the past few decades, various methods have been suggested by actuaries and demographers to project age-specific mortality rates. Pollard (1987) gives an excellent review of these methods. We only summarize methods adopted by actuaries in North America and the United Kingdom in the projection of future mortality rates.

1.1 The American Approach

The Society of Actuaries 1994 Group Annuity Reserving Table (GAR-94) [1] has adopted a generation life table approach to project mortality improvement.

Let q_x^z be the mortality rate observed at age x in calendar year z. Mortality improvement implies that the mortality rates for age x in future years form a non-increasing sequence in z. In the GAR-94 Table this implies that:

$$q_x^{1994} \geq q_x^{1995} \geq q_x^{1996} \geq q_x^{1997} \geq \cdots.$$

Let AA_x^z denote the annual improvement factor in the mortality rate for age x from calendar year z to $z+1$, i.e.,

$$AA_x^z = 1 - \frac{q_x^{z+1}}{q_x^z}.$$

The GAR-94 Table assumes that at each age the $AA_x^z = AA_x$, a constant, as z increases:

$$\frac{q_x^{1995}}{q_x^{1994}} = \frac{q_x^{1996}}{q_x^{1995}} = \frac{q_x^{1997}}{q_x^{1996}} = \cdots = 1 - AA_x. \tag{1}$$

To produce the mortality rate for a person age x in year $(1994+n)$, the following formula is used:

$$q_x^{1994+n} = q_x^{1994}(1 - AA_x)^n. \tag{2}$$

To assist in mortality projections using equation (2), the Society of Actuaries published the 1994 mortality rates as the base table, coupled with the improvement factors AA_x. Some values of q_x and AA_x for

[1]The 1994 Group Annuity Mortality Table and the 1994 Group Annuity Reserving Table are published by the Society of Actuaries Group Annuity Valuation Table Task Force in *Transactions of the Society of Actuaries*, 47 (1995): 865–913.

males at ages $x = 50, \cdots, 99$ are listed in Table 1 (where the values of q_x contain no margin). Specifically, Table 1 is an extract of the Society of Actuaries 1994 Group Annuity Reserving Table (the SOA GAR-94 Table) with: (i) base mortality rates q_x, (ii) improvement factors AA_x, and (iii) implied rate of improvement $\hat{E}_x$.

1.2 The British Approach

Based on the mortality experience in the United Kingdom, British actuaries have developed a more sophisticated method of projecting mortality.[2] Using the 1980 mortality rates as the base table, continuing improvement in mortality beyond 1980 is modeled as:

$$q_x^z = q_x^{1980}\{a(x) + [1 - a(x)](0.4)^{\frac{z-1980}{20}}\}, \tag{3}$$

where z is the calendar year, and

$$a(x) = \begin{cases} 0.5, & x < 60 \\ \frac{x-10}{100}, & 60 \le x \le 110 \\ 1 & x > 110. \end{cases} \tag{4}$$

Note that

$$\lim_{z \to \infty} q_x^z = a(x) q_x^{1980},$$

and

$$AA_x^z = 1 - \frac{a(x) + [1 - a(x)](0.4)^{\frac{z+1-1980}{20}}}{a(x) + [1 - a(x)](0.4)^{\frac{z-1980}{20}}} \tag{5}$$

is a decreasing function of z.

Equation (3) has three characteristics: (i) mortality improvement declines with advancing age; (ii) the mortality rate declines exponentially with the passage of time to a long-term limiting value; and (iii) the mortality improvement exhibits a decelerating trend.

---

[2] See Continuous Mortality Investigation Bureau (CMIB). "Standard Tables of Mortality Based on the 1979-82 Experiences." *Continuous Mortality Investigation Reports*, 10 (1990): 1-138.

Table 1
Excerpt of the Society of Actuaries'
Male 1994 Group Annuity Reserving Table (SOA GAR-94)

x	q_x	AA_x	$\hat{E}_x$	x	q_x	AA_x	$\hat{E}_x$
50	0.002773	0.018	0.01802	75	0.040012	0.014	0.01429
51	0.003088	0.019	0.01903	76	0.043933	0.014	0.01431
52	0.003455	0.020	0.02003	77	0.048570	0.013	0.01332
53	0.003854	0.020	0.02004	78	0.053991	0.012	0.01234
54	0.004278	0.020	0.02004	79	0.060066	0.011	0.01134
55	0.004758	0.019	0.01904	80	0.066696	0.010	0.01035
56	0.005322	0.018	0.01805	81	0.073780	0.009	0.009351
57	0.006001	0.017	0.01705	82	0.081217	0.008	0.008346
58	0.006774	0.016	0.01605	83	0.088721	0.008	0.008380
59	0.007623	0.016	0.01606	84	0.096358	0.007	0.007364
60	0.008576	0.016	0.01607	85	0.104559	0.007	0.007398
61	0.009663	0.015	0.01507	86	0.113755	0.007	0.007437
62	0.010911	0.015	0.01508	87	0.124377	0.006	0.006414
63	0.012335	0.014	0.01409	88	0.136537	0.005	0.005384
64	0.013914	0.014	0.01410	89	0.149949	0.005	0.005427
65	0.015629	0.014	0.01411	90	0.164442	0.004	0.004380
66	0.017462	0.013	0.01311	91	0.179849	0.004	0.004422
67	0.019391	0.013	0.01313	92	0.196001	0.003	0.003351
68	0.021354	0.014	0.01415	93	0.213325	0.003	0.003389
69	0.023364	0.014	0.01416	94	0.231936	0.003	0.003432
70	0.025516	0.015	0.01519	95	0.251189	0.002	0.002319
71	0.027905	0.015	0.01521	96	0.270441	0.002	0.002350
72	0.030625	0.015	0.01523	97	0.289048	0.002	0.002383
73	0.033549	0.015	0.01525	98	0.306750	0.001	0.001207
74	0.036614	0.015	0.01528	99	0.323976	0.001	0.001224

1.3 The Frailty Approach

Most actuarial and demographic techniques for projecting mortality rates are based on the extrapolation of past mortality rates. Few mathematical formulations are based on the underlying biological mechanism of mortality improvement.

Traditional life table methods, after accounting for factors such as race, gender, and smoking status, implicitly assume that the population is homogeneous, an assumption that is usually unrealistic. Empirical evidence shows that the following factors significantly affect mortality rates: genetics, economic status, education, marital status, and lifestyle. If mortality is not classified according to these additional risk factors, then the group's mortality characteristic will be heterogeneous.

For practical reasons not all of the above risk factors are usually included in mortality estimates. Thus, it is important to examine the consequences of heterogeneity when interpreting observed mortality rates and mortality improvements (Vaupel et al., 1979; Hougaard, 1991). A formal mathematical account of the treatment of heterogeneity can be found in Hougaard (1984, 1995) and the text of Namboodiri and Suchindran (1987).

Vaupel et al. (1979) propose a frailty model to study the effect of heterogeneity on cohort mortality rates.[3] In their model, each individual in a given population is endowed with his or her intrinsic *frailty index*, r, which is assumed to remain constant for life. An individual age x with frailty index r has force of mortality (hazard rate function), $\mu_x(r)$, which is assumed to satisfy

$$\mu_x(r) = r\mu(x) \tag{6}$$

where μ_x is the population's base force of mortality at age x. Weak (strong) individuals are associated with high (low) values of r.

2 Measurement of Mortality Improvement

To facilitate an easier discussion of mortality improvements, the following notational style is used:

[3]This frailty model can be viewed as a special version of the Cox (1972) proportional hazard model in the context of an unobserved covariate. Norberg (1989) uses a proportional hazard model for the heterogeneity in group life insurance. Two early actuarial applications of the frailty model that merit mentioning are Redington (1969) where there is a range of sample calculations, and Beard (1971) where the Gamma-Gompertz model is analysed.

- x denotes the current age and is placed at the right subscript;
- z denotes the current calendar year and is placed at the right superscript;
- A *bar* ($\bar{}$) placed on top of a quantity indicates that it is for a group of individuals; and
- A *hat* ($\hat{}$) placed on top of a quantity indicates that it is the estimated or observed value.

For example, $\hat{\mu}_x^z$ represents the estimated or observed hazard rate for an individual at age x, at calendar time z.

Customary measures of progress in mortality consider only changes in mortality rates $\bar{q}_x^z$ over different calendar years. Vaupel et al. (1979) argue that this may not be the most informative measure for mortality improvement. Instead of measuring progress in terms of mortality rates, Vaupel et al. state that it may be more appropriate to measure such progress in terms of the hazard rate (force of mortality) for standard individuals. Vaupel et al. (1979) give two main reasons:

1. For the frailty model of equation (6), the ratio of the μ's measures mortality progression at any level of frailty because the ratio is independent of r:

$$\frac{\mu_x^{z+n}(r)}{\mu_x^z(r)} = \frac{\mu_x^{z+n}(r')}{\mu_x^z(r')}.$$

However, this is not true for the ratio of the q's, i.e., the ratio depends on r:

$$\frac{q_x^{z+n}(r)}{q_x^z(r)} \neq \frac{q_x^{z+n}(r')}{q_x^z(r')}.$$

2. In youth and middle age, when μ_x and q_x are close to zero, μ_x is approximately equal to q_x. At the elderly ages, however, μ_x, which is not bounded by 1, can greatly exceed q_x. As a result, progress that substantially reduces μ_x may have much less effect on q_x. For example, consider a reduction in μ_x from 2 to 1: if these values of μ_x stayed constant over the course of a year, q_x would only be reduced from 0.86 to 0.63.

For an integer age x, we define annual improvement factors E_x^z in terms of hazard rates

$$E_x^z = 1 - \frac{\overline{\mu}_{x+0.5}^{z+1}}{\overline{\mu}_{x+0.5}^{z}} = 1 - \frac{\log(1 - \overline{q}_x^{z+1})}{\log(1 - \overline{q}_x^{z})} \tag{7}$$

where a constant hazard rate function is assumed for the age interval $(x, x + 1)$.

In the GAR-94 Table, the improvement factor AA_x is measured by the ratio of the observed mortality rates:

$$\hat{q}_x^{1994+n} = \hat{q}_x^{1994}(1 - AA_x)^n. \tag{8}$$

The implied improvement factor $\hat{E}_x^{1994}$ is

$$\hat{E}_x^{1994} = 1 - \frac{\log[1 - \hat{q}_x^{1994}(1 - AA_x)]}{\log[1 - \hat{q}_x^{1994}]}. \tag{9}$$

Table 1 shows the values of $\hat{E}_x^{1994}$ for comparison with the values of AA_x. From Table 1, one can see that the values of $\hat{E}_x^{1994}$ do not deviate much from AA_x for ages below 85. However, the relative difference becomes significant beyond age 85 and may affect our estimate of annuity costs (as they are based on mortality projections many years into the future).

3 A Mathematical Model for Frailty

3.1 The Basic Model

Consider a cohort of newborns (age exactly 0) where their survival capacity varies across individuals. A *standard* newborn is one whose future lifetime, X, has a force of mortality μ_x and cumulative force of mortality

$$H_x = \int_0^x \mu_t dt. \tag{10}$$

Each individual has his/her *unknown* constant frailty index r with force of mortality given in equation (6). Thus a standard newborn has $r = 1$.

To model the heterogeneity of frailty for the cohort of newborns age exactly 0, let R_0 be the unknown frailty index of an individual chosen at random from the cohort of newborns. Assume that R_0 has a probability density function (pdf) $g_0(r)$ for $r > 0$.

For a newborn with frailty r, the (conditional) survivor function and (conditional) pdf are:

$$\begin{aligned} \Pr[X > x | R_0 = r] &= S(x|r) = e^{-rH_x}, \\ f(x|r) &= -\frac{\partial}{\partial r} S(x|r) = r\, \mu_x\, e^{-rH_x}. \end{aligned}$$

The joint density of X and R_0 is

$$f(x,r) = f(x|r)\, g_0(r) = r\, \mu_x\, e^{-rH_x}\, g_0(r),$$

and the unconditional probability of a newborn chosen at random surviving to age x is

$$\begin{aligned} \Pr[X > x] &= \overline{S}(x) \\ &= \int_0^\infty S(x|r)\, g_0(r) dr \\ &= \int_0^\infty e^{-rH_x}\, g_0(r) dr \\ &= M_{g_0}(-H_x), \end{aligned} \tag{11}$$

where $M_{g_0}(\theta)$ is the moment generating function R_0, i.e.,

$$M_{g_0}(\theta) = E[e^{\theta R_0}] = \int_0^\infty e^{-r\theta}\, g_0(r) dr.$$

From $\overline{S}(x)$ we can get $\overline{f}(x)$, the pdf of X,

$$\overline{f}(x) = \mu_x \int_0^\infty r e^{-rH_x} g_0(r) dr, \tag{12}$$

and $\overline{\mu}_x$, the force of mortality associated with $\overline{S}(x)$,

$$\overline{\mu}_x = \frac{\overline{f}(x)}{\overline{S}(x)} = \frac{\mu_x \int_0^\infty r e^{-rH_x} g_0(r) dr}{M_{g_0}(-H_x)}. \tag{13}$$

Next we turn our attention to the survivors age exactly x from the cohort of newborns. Clearly the distribution of the frailty index among these survivors will not necessarily be the same as at age 0 because one would expect more of the weaker ones to have died earlier. So the population at age x should have a larger percentage of stronger individuals.

Let R_x be the frailty variable for the survivor cohort at age x chosen at random. R_x has pdf $g_x(r)$ given by

$$g_x(r) = \frac{S(x|r)g_0(r)}{\overline{S}(x)} = \frac{e^{-rH_x}\, g_0(r)}{M_{g_0}(-H_x)}. \tag{14}$$

Thus the average frailty for the survivor cohort at age x is

$$\overline{R}_x = E[R_x] = \frac{\int_0^\infty r e^{-rH_x} g_0(r)dr}{M_{g_0}(-H_x)} = \frac{\overline{\mu}_x}{\mu_x}. \tag{15}$$

Note that $\overline{\mu}_x = \mu_x \overline{R}_x$, i.e., the force of mortality for $[X > x]$ is always equal to the force of mortality of the standard individual multiplied by the average frailty among the survivors.

Among those who die at age x (i.e., in $(x, x + dx)$), the frailty index has a conditional density:

$$\frac{f(x,r)}{\overline{f}(x)} = \frac{r\, e^{-rH_x}\, g_0(r)}{\int_0^\infty r e^{-rH_x} g_0(r)dr}. \tag{16}$$

3.2 Gamma Frailty Density

Because of its mathematical tractability and its flexible shape, the gamma distribution has been used by many authors (including Vaupel et al., 1979) to model the frailty variable. Specifically we assume that R_0 has a gamma density:

$$g_0(r) = \frac{\beta^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r}, \quad r > 0 \tag{17}$$

where $\alpha > -1$ is a shape parameter and $\beta > 0$ is a scale parameter. The moment generating function is

$$M_{g_0}(\theta) = \left(\frac{\beta}{\beta - \theta}\right)^\alpha.$$

The first two moments of R_0 are:

$$\overline{R}_0 = E[R_0] = \frac{\alpha}{\beta}, \quad \text{and} \quad \sigma^2(R_0) = \frac{\alpha}{\beta^2}.$$

From equation (14) R_x is also gamma distributed with shape parameter α and a different scale parameter $\beta + H_x$.

In this case, the mean frailty of the survivors at age x is

$$\overline{R}_x = \frac{\alpha}{\beta + H_x}.$$

From equation (16), the frailty index for those who die in $(x, x + dx)$ has a conditional density that is also gamma distributed, with a shape parameter $\alpha + 1$ and a scale parameter $\beta + H_x$. In this case, the mean frailty of those who die is:

$$\overline{R}_x \times \frac{\alpha + 1}{\alpha},$$

which is greater than the mean frailty of the survivors.

3.3 Gompertz's Law

Assume that the standard individual's lifetime follows Gompertz's law:

$$\mu_x = bc^x \log(c), \qquad H_x = b(c^x - 1). \tag{18}$$

Gompertz's law has been used by actuaries since 1825. Several biological theories of aging have been developed that imply a Gompertz form of hazard rates (see Strehler, 1977, Chapter 5). Brillinger (1961) argues that if the human body is considered as a series system of independent components, then the hazard rate function may follow the Gompertz law (also see Carriere, 1992).

If the frailty variable R_0 is assumed to have a gamma density in equation (17), then the birth cohort has an unconditional survivor function

$$\overline{S}(x) = \left(\frac{\beta}{\beta + b(c^x - 1)}\right)^{\alpha},$$

and hazard rate function

$$\overline{\mu}_x = \frac{\alpha\, b\, c^x \log(c)}{\beta + b(c^x - 1)}. \tag{19}$$

Equation (19) is derived in the manner used by Beard (1971), and is one of the "laws" of mortality originally proposed by Perks (1932).

Pollard (1980, 1993) studies the case where each individual in a population has a hazard rate function of the Gompertz type. Note that the cohort hazard rate function in equation (19) increases exponentially initially, but the growth rate decreases with advancing age. Pollard points out that this is a phenomenon observed in many populations.

Among the survivors age x, R_x has a gamma distribution with a shape parameter α and a scale parameter $\beta + b(c^x - 1)$. The mean frailty for the survivors age x is

$$\overline{R}_x = \frac{\alpha}{\beta + b(c^x - 1)}.$$

4 A Model for Mortality Improvement

In a given calendar year, the overall level of mortality improvement depends on the marginal changes of many external factors such as medical technology and its availability to the general public. In general, projection of these external factors for future years is a difficult task and requires more detailed (perhaps non-actuarial) investigation. In this paper we are concerned mainly with the rates of mortality improvement among different cohorts in a given calendar year, where the same underlying external factors apply to all ages.

We hypothesize that mortality improvements due to the marginal advancement of life-saving techniques progress as follows:

Hypoyhesis 1. *For each age x the rate of improvement in terms of the force of mortality (hazard rate) is proportional to the average frailty $\overline{R}_x$.*

This hypothesis is based on the argument that marginal improvements in life-saving techniques have relatively larger effect on frailer individuals with higher than average values of r. Most deaths of strong individuals with lower than average values of r are due to natural aging; thus, improvements in life-saving techniques or better health-practices would have relatively smaller effects on healthier individuals, i.e., those with lower than average values of r.

Assume that, for each birth cohort at time y, frailty has a gamma density with:

$$g_0(r) = \frac{\beta^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r}, \qquad \text{where } \alpha = \beta.$$

If we assume Gompertz mortality for a standard individual, the mean frailty of the survivors at age x, at calendar time $z = y + x$, is

$$\overline{R}_x^z = \frac{\alpha}{\alpha + b(c^x - 1)}.$$

Based on the above hypothesis, the improvement factor E_x^z is proportional to the average frailty of the survivor cohort at age x:

$$E_x^z = \kappa \overline{R}_{x+0.5}^z = \kappa \frac{\alpha}{\alpha + b(c^{x+0.5} - 1)}, \tag{20}$$

where κ is a constant, and an adjustment of a half year is used because E_x^z is measured by the ratios of mid-year hazard rates.

Equation (20) of E_x^z implies that, at any fixed calendar time, the mortality improvement decreases rapidly at advanced ages, due to the exponential growth in $H_x = b(c^x - 1)$ with age.

5 Fitting the Gamma-Gompertz Model

Now we will fit the Gamma-Gompertz model to the GAR-94 Base Table, which gives the cohort age-specific mortality rates. Neither frailty nor heterogeneity was discussed in the GAR-94 Table. Specifically, we assume Gompertz's law for each individual's force of mortality.

Suppose that the cohort is homogeneous and each individual's lifetime follows the Gompertz law with $\mu_x = B c^x$. We would expect that

$$c[x] = [\frac{\mu_{x+20.5}}{\mu_{x+0.5}}]^{\frac{1}{20}}$$

be approximately constant.

From the GAR-94 mortality rates, we have calculated the values of $c[x]$ at different ages (see Table 2). As shown in Table 2, the values of $c[x]$ exhibit a gradual decreasing trend as age increases. Although there are many possible explanations to this observed pattern, we will try to fit the mortality rates to a frailty (heterogeneous) model.

Table 2
Calculated Values of $c[x]$

x	$c[x]$	x	$c[x]$
50	1.1180	60	1.1097
51	1.1171	61	1.1088
52	1.1160	62	1.1076
53	1.1151	63	1.1059
54	1.1142	64	1.1040
55	1.1133	65	1.1023
56	1.1124	66	1.1010
57	1.1114	67	1.1005
58	1.1107	68	1.1006
59	1.1102	69	1.1012

5.1 Fitting the 1994 Base Mortality Rates

Based on considerations that mortality rates at advanced ages may not be as accurate due to smaller sample sizes, we suggest using some representative age range, say, from 50 to 75. For many populations, from the mortality rates at ages 50 and 70, one can get a good approximation of the shape of the mortality curve at all ages (Benjamin, 1982; Pollard, 1991 and 1993).

We assume that the standard force of mortality follows the Gompertz law with

$$\mu_x = bc^x \log(c), \qquad H_x = b(c^x - 1).$$

Furthermore, we will choose μ_x such that $\bar{R}_0 = 1$.

We define a measure for goodness of fit by using the sum of squared errors for ages from 50 to 75:

$$\text{DIST}_{50:75} = \sum_{x=50}^{75} (\overline{\mu}_{x+0.5} - \hat{\mu}_{x+0.5})^2$$

where

$$\hat{\mu}_{x+0.5} = -\log(1 - q_x)$$

can be obtained from the mortality rates in Table 1.

Assume that R_0 has a gamma density, then $\bar{R}_0 = 1$ implies $\alpha = \beta$ so that

$$\overline{\mu}_x = \frac{\alpha\, b\, c^x \log(c)}{\alpha + b(c^x - 1)}. \tag{21}$$

By minimizing $\mathrm{DIST}_{50:75}$, we get the following estimate of the unknown parameters:

$$c = 1.1248, \qquad b = 0.66 \times 10^{-4}, \qquad \alpha = 1.306, \tag{22}$$

with the minimum distance being

$$\min\{\mathrm{DIST}_{50:75}\} = 0.2238 \times 10^{-5}.$$

We have noticed that the estimation of parameters for the frailty distribution (mixing density) is not very robust, depending on the age range used in the estimation of parameters. This is a common phenomenon in many mixture models (Chan, 1995; Manton et al., 1986; Everitt and Hand, 1981).

5.2 Fitting the GAR-94 Mortality Improvement Factors

We shall use a Gamma-Gompertz model for the 1994 base mortality rates and adopt the particular set of estimated parameters in equation (22) in the Gamma-Gompertz model:

$$c = 1.1248, \qquad b = 0.66 \times 10^{-4}, \qquad \alpha = 1.306.$$

The frailty model of mortality improvement in equation (20) suggests the following pattern for the improvement factors:

$$E_x^{1994} = \kappa \times \frac{1.306}{1.306 + 0.66 \times 10^{-4} \times (1.1248^{x+0.5} - 1)}.$$

Now we use this frailty model of mortality improvement to fit the empirical improvement factors $\hat{E}_x^{1994}$ in Table 1. We are mainly interested in the mortality improvement at senior ages, say, 50 and above. We first choose an age range from 50 through 95 and define a loss measure:

$$M_{50:95} = \sum_{x=50}^{95} (E_x^{1994} - \hat{E}_x^{1994})^2.$$

The ages below 50 are excluded because of the sudden dip in the observed improvement factors (see Figure 1) which may be a result of other exogenous factors (e.g., accident, AIDS). The ages beyond 95 are not included because of the scarcity of available data for extreme ages 95 and above.

By minimizing the loss measure $M_{50:95}$, we get a least square estimate for κ:

$$\kappa = \frac{\sum_{50}^{95}(E_x^{1994})(\hat{E}_x^{1994})}{\sum_{50}^{95}(E_x^{1994})^2} = 0.01769.$$

Table 3 compares the Gamma-Gompertz frailty model improvement factors E_x and the empirical improvement factors $\hat{E}_x$ in the GAR-94 table. Figure 1 also displays these improvement factors. Note that in Figure 1 the $\hat{E}_x$s in the GAR-94 Table do not follow a smooth pattern. Also, there is insignificant mortality improvement in the 25-45 age group.[4] Beyond age 50 the frailty model seems to be an acceptable fit and may provide a theoretical basis for the observed improvement factors. The frailty model of mortality improvement has the definite advantage that the projected mortality rates are smooth.

The choices of the age range, from 50 to 95, and the loss measure (i.e., the squared error) are arbitrary and are for illustration purposes only. One may use other age ranges or weighted squared error, as appropriate.

6 Other Evidence

According to United Nations 1991,[5] in developed countries, one half of female and one-third of male deaths now occur after age 80. The mortality reductions within this age range are crucial in determining changes in life expectancy and actuarial annuity values.

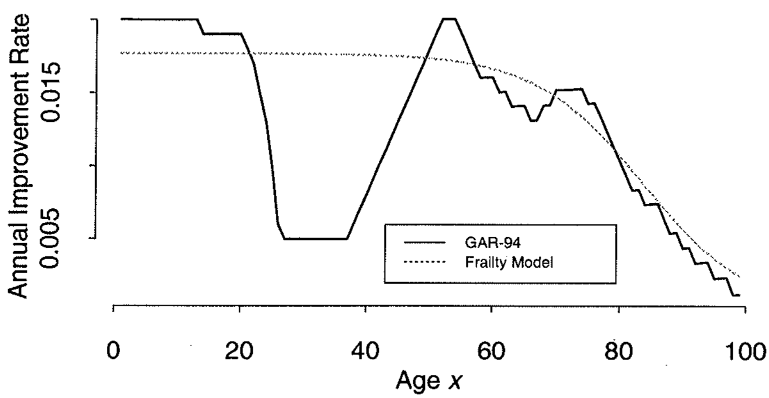
[4]This could be attributed to extra AIDS deaths in the 25-45 age group.

[5]United Nations Demographic Yearbook 1991. New York: United Nations

Table 3
Mortality Improvement Rates

x	E_x	$\hat{E}_x$	x	E_x	$\hat{E}_x$
50	0.017410	0.018030	75	0.013520	0.014290
51	0.017370	0.019030	76	0.013130	0.014320
52	0.017330	0.020030	77	0.012720	0.013330
53	0.017290	0.020040	78	0.012290	0.012340
54	0.017240	0.020040	79	0.011840	0.011340
55	0.017180	0.019050	80	0.011370	0.010350
56	0.017120	0.018050	81	0.010890	0.0093500
57	0.017050	0.017050	82	0.010390	0.0083500
58	0.016980	0.016050	83	0.0098800	0.0083800
59	0.016890	0.016060	84	0.0093700	0.0073600
60	0.016800	0.016070	85	0.0088500	0.0074000
61	0.016690	0.015070	86	0.0083300	0.0074400
62	0.016580	0.015080	87	0.0078100	0.0064100
63	0.016450	0.014090	88	0.0073100	0.0053800
64	0.016310	0.014100	89	0.0068100	0.0054300
65	0.016150	0.014110	90	0.0063200	0.0043800
66	0.015980	0.013110	91	0.0058500	0.0044200
67	0.015790	0.013130	92	0.0054000	0.0033500
68	0.015580	0.014150	93	0.0049700	0.0033900
69	0.015350	0.014160	94	0.0045600	0.0034300
70	0.015100	0.015190	95	0.0041800	0.0023200
71	0.014830	0.015210	96	0.0038200	0.0023500
72	0.014530	0.015230	97	0.0034800	0.0023800
73	0.014220	0.015260	98	0.0031600	0.0012100
74	0.013880	0.015280	99	0.0028700	0.0012200

Figure 1
Mortality Improvement Factors $\hat{E}_x$ for GAR-94 And E_x for the Frailty Model



Kannisto et al., (1994) study the reduction in mortality at advanced ages based on a large and reliable database for 27 countries, 1960s through 1980s. The following is cited from Kannisto et al., (1994, pp. 801):

> For nine countries - Austria, Belgium, England and Wales, West Germany, France, Japan, Scotland, Sweden, and Switzerland - data are available through 1991. A glimpse at the most recent trends is provided by calculating the annual average rate of mortality improvement between 1982-86 and 1987-91 for this aggregate of nine countries. For males the rate of improvement was 1.7 percent for octogenarians and 1.2 percent for nonagenarians; for females the corresponding rates were 2.5 percent and 1.6 percent.

Even though the magnitude of the mortality improvement at advanced ages is higher than those in the GAR-94 Table, the general pattern of deceleration of mortality improvement at advanced ages is consistent with our frailty model of mortality improvement.

In a panel discussion of mortality trends, Moriyama (1967) also provides evidence that the rate of improvement in mortality rates decreases at advanced ages.

7 Closing Comments

The main contribution of this paper is the utilization of a frailty model to derive mathematical formulae for mortality improvement factors. As marginal advancement in life-saving techniques determines the pace of mortality improvement, we assume that weaker individuals are more likely to benefit from these advances than are stronger individuals. This assumption is supported in the demography literature (Vaupel and Yashin, 1985). To project the future trend of mortality improvement, one needs to assess carefully the future advancement in medical technology. A major breakthrough in medical technology or an unexpected new epidemic may have a sudden impact on the mortality improvement.

Several authors, including Bowers et al., (1986) and London (1985), have discussed the importance of smoothness in mortality rates. Their arguments for smoothness can be extended to mortality improvement factors. Our frailty model provides useful mathematical formulae for graduation of empirical improvement factors.

One potential shortcoming of our model is that the frailty index is assumed to be determined at birth and remains constant for life. Intuition suggests, however, that this assumption may be overly simplistic. In future studies, the concept of frailty may be modeled as a variable dependent upon exogenous observable factors such as lifestyle, environment, economic status, or marital status.

We hope this paper stimulates further research on this important subject.

References

Beard, R.E. (1971). "Some Aspects of Theories of Mortality, Cause of Death Analysis, Forecasting and Stochastic Processes." In *Biological Aspects of Demography* (ed W. Brass). London, England: Taylor and Francis, 1971: 57-68.

Benjamin, B. "The Span of Life." *Journal of the Institute of Actuaries* 109 (1982): 319-340.

Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A. and Nesbitt, C.J. *Actuarial Mathematics.* Itasca, Ill.: Society of Actuaries, 1986.

Brillinger, D.R. "A Justification of Some Common Laws of Mortality." *Transactions of the Society of Actuaries* 13, part 1, (1961): 116-126.

Carriere, J.F. "Parametric Models for Life Tables." *Transactions of the Society of Actuaries* 44 (1992): 77-99.

Chan, J. "Optimal Rate of Convergence for Finite Mixture Models." *The Annals of Statistics* 23, no. 1, (1995): 221-233.

Cox, D.R. "Regression Models and Life Tables." *Journal of the Royal Statistical Society,* Series B 34 (1972): 187-202.

Everitt, B.S. and Hand, D.J. *Finite Mixture Distributions,* London: Chapman and Hall Ltd, 1981.

Gompertz, B. "On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies." *Philosophical Transactions of the Royal Society,* 115 (1825): 513-85.

Hougaard, P. "Life Table Methods for Heterogeneous Populations: Distributions Describing the Heterogeneity." *Biometrika* 71, no. 1 (1984): 75-83.

Hougaard, P. "Modeling Heterogeneity in Survival Data." *Journal of Applied Probability* 28 (1991): 695-701.

Hougaard, P. "Frailty Models for Survival Data." *Lifetime Data Analysis* 1, no. 3 (1995): 255-273.

Jenkins, W.A. and Lew, E.A. "A New Mortality Basis for Annuities." *Transactions, Society of Actuaries* 1 (1949): 369-466.

Kannisto, V., Lauritsen, J., Thatcher, A.R. and Vaupel, J.W. "Reductions in Mortality at Advanced Ages: Decades of Evidence from 27 Countries." *Population And Development Review* 20, no. 4 (1994): 793-810.

Lancaster, H.O. *Expectations of Life: A Study in the Demography, Statistics, and History of World Mortality.* New York, N.Y.: Springer-Verlag, 1990.

London, D. *Graduation: The Revision of Estimates.* Winsted, Conn.: ACTEX Publications, 1985.

Manton, K.G., Stallard, E. and Vaupel, J.W. "Alternative Models for the Heterogeneity of Mortality Risks Among the Aged." *Journal of the American Statistical Association* 81, no. 395, (1986): 635-644.

Moriyama, I.M. Panel discussion on "Mortality Trends and Projections." *Transactions, Society of Actuaries* 19, part 2 (1967): D429-493.

Namboodiri, K. and Suchindran, C.M. *Life Table Techniques and Their Applications.* New York, N.Y.: Academic Press, 1987.

Norberg, R. "Experience Rating in Group Life Insurance." *Scandinavian Actuarial Journal* (1989): 194-224.

Perks, W. (1932). "On Some Experiments in the Graduation of Mortality Statistics." *Journal of Institute of Actuaries* 63 (1932): 12-40.

Pollard, J.H. "Methodological Issues in the Measurement of Inequality of Death." In *Mortality in South and East Asia—A Review of Changing Trends and Patterns, 1950-1975.* World Health Organization, 1980.

Pollard, J.H. "Projection of Age-Specific Mortality Rates." *Population Bulletin of the UN* no. 21/22 (1987): 55-69.

Pollard, J.H. "Fun with Gompertz." *Genus* 47 (1991): 1-20.

Pollard, J.H. "Heterogeneity, Dependence Among Causes of Death and Gompertz." *Mathematical Population Studies* 4, no. 2, (1993): 117-132.

Redington, F.M. "An Exploration into Patterns of Mortality." *Journal of Institute of Actuaries* 95 (1969): 243-298.

Strehler, B.L. *Time, Cells, and Aging.* New York, N.Y.: Academic Press, 1977.

Vaupel, J.W., Manton, K.G. and Stallard, E. "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality." *Demography* 16, no. 3 (1979): 439-454.

Vaupel, J.W. and Yashin, A.I. "Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics." *Journal of the American Statistical Association* 39, no. 3 (1985): 176-185.

Vaupel, J.W. and Yashin, A.I. "Repeated Resuscitation: How Lifesaving Alters Life Tables." *Demography* 4, no. 1 (1987): 123-135.